A MONTE CARLO INVESTIGATION OF THREE DIFFERENT ESTIMATION METHODS IN MULTILEVEL STRUCTURAL EQUATION MODELING UNDER CONDITIONS OF DATA NONNORMALITY AND VARIED SAMPLE SIZES

A Dissertation

by

JIMMY KENT BYRD

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree
of

DOCTOR OF PHILOSOPHY

December 2008

Major Subject: Educational Psychology

A MONTE CARLO INVESTIGATION OF THREE DIFFERENT ESTIMATION METHODS IN MULTILEVEL STRUCTURAL EQUATION MODELING UNDER CONDITIONS OF DATA NONNORMALITY AND VARIED SAMPLE SIZES

A Dissertation

by

JIMMY KENT BYRD

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree
of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee, Bruce Thompson
Committee Members, Victor Willson
Oi-Man Kwok
Robert Capraro
Head of Department, Victor Willson

December 2008

Major Subject: Educational Psychology

ABSTRACT

A Monte Carlo Investigation of Three Different Estimation Methods in Multilevel Structural Equation Modeling Under Conditions of Data Nonnormality and Varied Sample Sizes. (December 2008)

Jimmy Kent Byrd, B.S., Tarleton State University; M.Ed., Tarleton State University; Ph.D., Texas A&M University

Chair of Advisory Committee: Dr. Bruce Thompson

The purpose of the study was to examine multilevel regression models in the context of multilevel structural equation modeling (SEM) in terms of accuracy of parameter estimates, standard errors, and fit indices in normal and nonnormal data under various sample sizes and differing estimators (maximum likelihood, generalized least squares, and weighted least squares). The finding revealed that the regression coefficients were estimated with little to no bias among the study design conditions investigated. However, the number of clusters (group level) appeared to have the greatest impact on bias among the parameter estimate standard errors at both level-1 and level-2. In small sample sizes (i.e., 300 and 500) the standard errors were negatively biased. When the number of clusters was 30

and cluster size was held at 10, the level-1 standard errors were biased downward by approximately 20% for the maximum likelihood and generalized least squares estimators, while the weighted least squares estimator produced level-1 standard errors that were negatively biased by 25%. Regarding the level-2 standard errors, the level-2 standard errors were biased downward by approximately 24% in nonnormal data, especially when the correlation among variables was fixed at .5 and kurtosis was held constant at 7. In this same setting (30 clusters with cluster size fixed at 10), when kurtosis was fixed at 4 and the correlation among variables was held at .7, both the maximum likelihood and generalized least squares estimators resulted in standard errors that were biased downward by approximately 11%.

Regarding fit statistics, negative bias was noted among each of the fit indices investigated when the number of clusters ranged from 30 to 50 and cluster size was fixed at 10. The least amount of bias was associated with the maximum likelihood estimator in each of the data normality conditions examined. As sample size increased, bias decreased to near zero when the sample size was equal to or greater than 1,500 with similar results reported across estimation methods. Recommendations for the substantive researcher are presented and areas of future research are presented.

ACKNOWLEDGEMENTS

This dissertation is dedicated to my daughter, Kenzie Sue Byrd. Kenzie has always shown unconditional and enduring love even in times when I could not be there for her during my completion of the doctoral program.

Regarding acknowledgements, the doctoral program in Educational Psychology at Texas A&M University has been such a blessing to me. I was fortunate to be a part of a program that combines rigor and scholarly learning in an atmosphere of excellence. The professors at TAMU strongly exemplify servant leadership and collegiality. I attended classes with some of the finest Christian people, who challenged and questioned my philosophies and beliefs, yet in a supportive manner. I would like to thank my dissertation committee especially Bruce Thompson for chairing my committee and providing insight and assistance along my journey by challenging me to pursue quality research with sound statistical measures.

In addition, I would like to thank Victor Willson who sparked my interest in structural equation modeling and for always being available to answer questions and provide guidance.

Most importantly, I acknowledge my marvelous parents, Kay and Woody Byrd. They have been the most influential people in my life. I am grateful for the encouragement and support that they have provided throughout my life.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

TABLE Page

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Multilevel modeling refers to a variety of statistical methods that are utilized to examine data that maintain a nested or hierarchical structure (NCES, 2004). This includes data in most educational settings where students are nested within classrooms and classrooms are nested within schools. Many national surveys use a complex sampling design that includes clustered sampling (NCES). Only recently, according to Draper (1995), has the multilevel nature of sampling designs been considered as a technique to analyze such data. Indeed, multilevel modeling is an important tool to utilize as standard analytic methods are biased by deviations from simple random sampling (Bryk & Raudenbush, 1992). In clustered sampling, sizeable pockets of similarity among the individuals comprising each group when data are nested are highly probable, negating the standard statistical tests, which depend heavily on the assumptions of independence (Hox, 1995).

This dissertation follows the style of *The Journal of Educational Research*.

Hierarchical Linear Modeling (HLM) has been widely applied in educational and behavioral sciences as the currently-preferred analysis for multilevel data. This preferred use has been based on assumptions of improved estimation of standard errors of individual effects and appropriate partitioning of variance-covariance components under a random sampling model. HLM is capable of fitting random effect models, contextual effect models and cross-level interaction models (Hox & Kreft, 1994). Being in the general framework of multivariate statistics, HLM maintains the traditional assumptions of linear model analysis for linearity and normality, but allows violation of homoscedasticity and more importantly, observation independence (Raudenbush & Bryk, 2002).

A variety of statistical programs that can be used for conducting a multilevel analysis fall primarily into two categories: (a) multilevel regression programs that focus on single outcomes and the investigation of random intercepts and slopes; and (b) multilevel covariance structure analysis that focuses on relations among latent variables (Heck & Thomas, 2000). Both approaches, multilevel regression and multilevel covariance structure analysis, have a close connection in that multilevel regression can theoretically be specified as a special case of the more general structural equation model as a linear model whose covariance structure can be written in finite form (Bollen 1989; Hox, 2002). Despite similarities, recent interest in conducting multilevel regression analysis in the covariance structure analysis or structural equation modeling (SEM) framework has developed. Current SEM software such as LISREL (Jöreskog & Sörbom, 1996), EQS (Bentler, 1995), AMOS (Arbuckle & Wothke, 1999), and M-Plus (Muthén & Muthén, 1998-2004) can be adapted to implement multilevel SEM.

SEM is rapidly gaining popularity in the social science (Curran, 2003; Tomarken & Baker, 2003). A plausible explanation for the increased use of SEM is that user-friendly software packages have made SEM accessible to substantive researchers who may have limited training in statistics (e.g., Arbuckle & Wothke, 1999; Bentler, 1995; Jöreskog & Sörbom, 1996; Jöreskog, Sörbom, Du Toit, & Du Toit, 2000; Muthén & Muthén, 2002; Steiger, 1995).

Multilevel regression modeling (MRM) is compared with multilevel structural equation modeling (MSEM) by Tomarken and Waller (2005). Tomarken and Waller noted the advantages of MSEM to be more interpretable measures of goodness of fit, better modeling of residuals, and better capacity to model latent variables. The researchers also cited the advantages of MRM to be easier model specification, fewer estimation problems, and ability to handle certain types of analysis difficult to handle within MSEM. For instance, in MSEM one cannot model between-group variability in factor pattern or path coefficients (Hox, 2002). Tomarken and Waller surmised that the two approaches are more similar than different.

Multilevel modeling is an attractive approach to the social sciences as it allows the incorporation of substantive theory about organizational effects at different levels within the clustered nature of the data. Despite the attractiveness of multilevel modeling, a variety of issues must be resolved when applying multilevel models in the SEM framework of organizational research. These issues include determining which estimator should be used when data exhibit departures from normality and how varying sample size impacts parameter and standard error estimates, as well as fit indices, when employing various estimators in the event data are nonnormal (Bickel,2007).

Both HLM and SEM are multivariate statistics that require multivariate normally distributed data. However, in real-world situations, the assumption of multivariate normality is not often met (Heck & Thomas, 2000; Micerri, 1989; Zhang, 2005). According to Zhang, ignoring nonnormality and treating data as normal will result in biased standard errors and parameter estimates, leading to misleading interpretations, which could be detrimental in certain situations.

The importance of adhering to assumptions and selecting appropriate strategies and estimators to model data based on the characteristics of the data cannot be ignored. As mentioned earlier, violations of the assumptions of SEM can lead to biased results that may lead to incorrect decisions about the theory or model being tested (Finely & DiStefano, 2006). As with most statistics and particularly SEM, certain assumptions must be met in order for researchers to trust the obtained results. A primary consideration in SEM is the choice of the estimation methods used to obtain parameter values, standard errors, and fit indices (Brown, 2006). Generally, the effects of non-normality on normal theory estimators (Maximum Likelihood & Generalized Least Squares) depend on the extent of non-normality (Finney & DiStefano, 2006). Furthermore, increasing non-normality leads to a greater impact on the obtained results. Consequently, it is important that researchers examine the distribution of the observed variables prior to beginning the SEM analysis.

When data do not meet the assumptions for normal theory estimators, namely observed endogenous variables exhibit departure from normality, observations are not independent, or the data are categorical in nature, one may employ an asymptotic distribution free (ADF) estimator such as weighted least squares (WLS). The WLS method has several advantages as well as some disadvantages (Bollen, 1989). One main advantage is that WLS requires only minimal assumptions about the distribution of the observed variables. Simulation research with nonnormal data shows that the WLS test statistic is relatively unaffected by distributional characteristics. However, one of the primary disadvantages of the WLS method compared to ML is that WLS requires large samples in order to obtain consistent and efficient estimates. If data are continuous but nonnormal, the estimation method most often recommended is the asymptotically distribution free (ADF) method (Browne, 1984). Although simulation studies suggest that maximum likelihood estimation, with or without a correction for nonnormality, seems to perform better than ADF and should be preferred (Boomsma & Hoogland, 2001; Hu, Bentler, & Kano, 1992).

Olsson, Foss and Howell (2000) studied the effects of three estimation methods (ML, GLS, & WLS) on indexes of fit and parameter bias for different sample sizes when nested models vary in terms of specification error and differing levels of kurtosis. The results revealed that maximum likelihood compared to GLS under conditions of misspecification providing more realistic indexes of overall fit and less bias in parameter estimates for paths that overlapped the true model. Further, Olsson, Foss and Howell suggested that WLS was not preferable to either ML or GLS in terms of parameter bias or model fit even though with large samples sizes and mildly misspecified models, WLS provided estimates close to the ML and GLS estimators.

However, each of the studies reviewed were limited to factor analysis and not multilevel regression models within the SEM framework.

Recently, Zhang (2005) compared both HLM and SEM in terms of performance with multiple imputed data with data missing at random and missing completely at random with model misspecification using maximum likelihood estimation. Zhang found that HLM and SEM had similar power to detect the main effect of first level predictor and cross-level interaction effect across various sample sizes, and higher power was found with larger samples. However, data with small sample size may not be appropriate for multilevel analyses because there was lack of sufficient information at either level for stable estimation. Under severely nonnormal data conditions, HLM had better power for the interaction effect than SEM, but neither method was very sensitive to violations of normality in terms of parameter estimates and standard errors. Although Zhang's study employed maximum likelihood estimation, it was not clear how the weighted least squares (WLS) or generalized least squares (GLS) estimators would affect parameter estimates, standard error restimates, and fit indices when the data exhibit departure from normality or sample and cluster size varied.

Significance of the Study

Multilevel regression modeling in the SEM framework is becoming increasingly popular in the social sciences. The recent upsurge of interest in SEM does, however, have a potential downside. As Steiger (2001) has pointed out, an unfortunate consequence of SEM's increased popularity and accessibility may be lack of awareness on the part of many users of SEM's assumptions, limitations, complexities, and ambiguities. Given the frequency of SEM's use, researchers must recognize the assumptions associated with different estimation methods to ensure that the parameter estimates, standard errors and fit indices are correct to ensure, in turn, proper interpretation of results.

Purpose and Research Questions

The purpose of the present study was to examine a multilevel regression model in the context of multilevel SEM in terms of accuracy of parameter estimates, standard errors, and fit indices with nonnormal data and model misspecification under various sample and cluster sizes with differing estimators (ML, GLS, and WLS). Due to the number of factors examined, balanced sample size was maintained. The specific research questions addressed in this study included:

1. Do the parameter estimates differ by the design factors (estimation methods, nonnormality conditions, and sample and cluster size)?
2. Do standard errors differ by the design factors (estimation methods, nonnormality conditions, and sample and cluster size)?
3. Do fit indices differ by the design factors (estimation methods, nonnormality conditions, and sample and cluster size)?

Delimitations of the Study

This study was delimited in the number of combinations of sample size, nonnormal data conditions and estimators employed. Further, the models examined are simple regression models. Cautious interpretation of the results in terms of more complex models is warranted.

CHAPTER II

REVIEW OF THE LITERATURE

This literature review provides information about studies related to multilevel modeling of nonnormal data. The chapter is divided into six distinct sections: a brief and general introduction of multilevel modeling using hierarchical linear modeling (HLM) and structural equation modeling (SEM); a review of empirical research studies investigating sample and cluster sizes; a review of the intraclass coefficients and the design effect; a review of commonly applied fit indices; a comprehensive review of research studies dealing with nonnormal data in a multilevel context and a review of estimation methods available in the SEM framework. All six sections are intended to provide readers an overview of the research problem and the findings of previous literature that influenced the hypotheses of the current study.

Multilevel Modeling Using HLM and SEM

Although maximum likelihood is the most common estimator employed when conducting SEM analyses with normally distributed data, others have examined estimation of parameters and fit statistics in nonnormal distributions under the weighted least squares (WLS) estimation methods in factor analysis (Olsson, Foss, Troye & Howell, 2000). However, little is known about the performance of parameter estimates, standard errors and fit statistics in multilevel regression conducted in the SEM framework when the outcome variable exhibits departure from normality and/or the model is misspecified.

With the advent of the expectation-maximization (EM) algorithm using maximum likelihood estimation developed by Dempster, Laird, and Rubin (1977), multilevel modeling has become a primary mode of analyzing data that are nested or maintain a hierarchical structure. The nesting of observations within groups, such as students clustered within classrooms or schools, is fundamental to multilevel modeling. Indeed, nested or hierarchically structured data should be of primary importance in that such data incorporate contextual factors, rather than focusing exclusively on individual-level variables (Bickel, 2007; Bryk & Raudenbush, 1992). According to Zhang (2005), most traditional statistical procedures disaggregate data from higher levels into individual-level variables and treat data as though the data were individual-level variables, or vice versa. Dealing with multilevel structured data in an individual-level manner omits important information such as the independence of responses and within-cluster information (Goldstein, Browne, & Rasbash, 2002; Hox, 1995). In nested data, there are likely nested pockets of data that are very similar within group clusters. Using standard individual-level models will result in unbiased but inefficient parameter estimates and underestimate the standard errors of the model parameters, leading to increased Type I errors (Heck & Thomas, 2007; Mass & Hox, 2004).

Hierarchical linear models (Bryk & Raudenbush, 2002), also known as random coefficient models (Kreft & de Leeuw, 1998), variance component models (Longford, 1989), or multilevel random coefficient models (Nezlek, 2001, 2003), take the multilevel structure of data into account. Hierarchical linear models were originally developed in educational and social research where observations are often made on different levels simultaneously (e.g., students, classes, schools).

An important feature of HLM is its capacity to study cross-level effects, which usually focus on the contextual effect of a higher-level variable on lower level effects. HLM is particularly attractive because in social sciences research the primary main effect of interest is not always statistically significant but still may interact with certain contextual characteristics (Zhang, 2005).

The assumptions underlying the multilevel regression model are similar to the assumptions in ordinary least squares regression. Namely, the assumptions include linear relationships, homoscedasticity, and normal distribution of the residuals. In multiple regression, given a large sample size, it is widely known that a violation of these assumptions does not lead to highly inaccurate parameter estimates or standard errors (Maas & Hox, 2004a). In the case of extreme violations, many statistical methods are available for correcting homoscedasticity (Scott, Long, & Ervin, 2000). However, in multilevel regression analysis, heteroscedasticity can be modeled directly (Goldstein, 1995). More concretely, the assumptions underlying the multilevel model include linear relations, a normal distribution for the individual-level residuals, and a multivariate normal distribution for the group-level residuals. Note the residuals from the individual-level are assumed independent from the group-level residuals.

Sample and Cluster Size

A common estimator used in the multilevel modeling framework includes asymptotic maximum likelihood methods, which are based on the assumption of large sample size (Eliason, 1993). One important assumption underlying the maximum likelihood estimation method is normality of the error distributions. When the residual errors are not normally distributed, the parameter estimates produced by the maximum likelihood method are consistent and asymptotically unbiased. However, the asymptotic standard errors are not correct.

While research in educational settings is generally based on varying sample sizes, the question is how do asymptotic maximum methods perform in small sample sizes? This is especially important considering that the sample size at level-2 is by definition smaller than the level-1 sample size (Mass & Hox, 2004b). In simulation studies Van Der Leeden and Busing (1994), suggested that when the assumption of large sample size and normality do not hold, the group level variance components tend to be underestimated while the standard errors exhibit a downward bias. As a result, it was reported that for highly accurate estimates of group level variance components, more than 100 groups are needed. In contrast, Browne and Draper (2000) suggested that with as few as six to twelve groups, restricted maximum likelihood (RML) estimation provides reasonable variance estimates and, with 48 groups, both RML and full information maximum likelihood (FML) estimation produces reasonable variance estimates. However, Van Der Leeden, Busing, and Meijer (1997) reported that the standard errors of the variance components are generally too small when estimated with FML, while RML was most accurate.

Similarly, in a large simulation study employing RML, Mass and Hox (2005) found that standard errors for the regression coefficients were slightly biased downward if the number of groups was less than 50. Regarding the individual sample size, in Monte Carlo simulation studies, Willson and Zhang (2003) and Zhang and Willson (2006) reported that the first level sample size did matter. In a two-level cross-interaction model, it was found that power did not increase further when the group size exceeded 35, when the number of groups was fixed at 120. While the results are not in total agreement, the underlying message is that the regression coefficients are unbiased while standard errors are biased downward with small sample size at the group level when employing the maximum likelihood estimator. Thus, in these situations significance tests and confidence intervals cannot be fully trusted. According to Goldstein (1995), the problem does not completely vanish as the sample size increases.

Based on prior simulation studies related to sample size at level-one and the number of clusters or groups at level-two, there are recommendations or ‘rules of thumb’ that have been reported to lead to stable results. Kreft (1996) suggested a ‘30/30’ rule of thumb. According to Kreft, researchers should strive for a sample of at least 30 groups with at least 30 individuals per group if interest is primarily in the fixed effects. Hox (n.d.) submitted that sample size should be increased to 50/20 (50 groups with a minimum of 20 individuals per group) if interest is on the cross-level interactions. Further, if interest lies in the random portion of the model, Hox suggested that the variance and covariance should be based on a 100/10 rule (100 groups with a minimum of 10 individuals per group).

Intraclass Coefficient and Design Effect

Hox and Maas (2001) reported that the size of the intraclass correlation (ICC) also impacted the accuracy of the estimates. According to Hox and Maas, in general, what is at issue in multilevel modeling is not so much the ICC but the design effect (DEFF), which indicates how much the standard errors are underestimated (p. 431). As Kish (1965) demonstrated, in cluster samples, the design effect is approximately equal to (1 + [(average cluster size - 1) * ICC)]). Employing DEFF, the standard statistical formulas can be adjusted to reflect the true sampling variance. If such adjustments are made, the impact of cluster sampling on the operating alpha can be rather large. For example, Barcikowski (1981) examined the effect of cluster sampling on the actual alpha level of a *t*-test conducted at the nominal alpha level of .05. With an intraclass correlation of rho = .05 and an average cluster size of 10, the operating alpha level is 0.11. With large intraclass correlations and larger cluster sizes, the operating alpha level increases rapidly. However, if the design effect is smaller than two, using single-level analysis on multilevel data does not lead to overly misleading results (Muthen & Satorra, 1995).

Multilevel Models in the SEM Framework

While there are many computer programs specifically designed to model multilevel data such as HLM (Bryk , Raudenbush, Seltzer,& Condon, 1988), Proc Mixed in SAS (SAS Institute, 2002), MLwiN (Rasbash, et al., 2000) the mixed modeling package in SPSS (SPSS, 2005), and Mplus (Muthen & Muthen, 1998-2006), recent interest has focused modeling multilevel data in a structural equation framework (Bauer, 2003; Bentler & Liang, 2003; Curran, 2003; Mehta & Neale, 2005). Multilevel modeling and SEM have evolved from different conceptual and methodological roots. Multilevel modeling focuses on clustered data and attempts to partition the observed variance into within- and between-clusters components. SEM, on the other hand, focuses on modeling means and covariances among multivariate data (Bollen, 1989). Because both clustered and multivariate data are inherent in educational and psychological research, it is not surprising that practitioners of each method are interested in borrowing the strengths of the two techniques (Goldstein & McDonald, 1988; Mehta & Neale, 2005).

Although the intersection of multilevel models and SEM is being analyzed in various current settings, this notion is not new. Meredith and Tisak (1984, 1990) were some of the first researchers to fit a multilevel model in the SEM framework. Another pair of researchers, McArdle and Hamagami (1996), used a multigroup SEM analysis to estimate subsets of multilevel models. Further, Rovine and Moelnar (1998, 2000, 2001) explored the intersection between SEM and multilevel models using separate structures for the fixed and random effect to stay maximally consistent with the Laird and Ware (1982) expressions. More recently, Newsome (2002) explored a creative variant of these models to use the SEM framework to estimate a multilevel model for dyadic data. Even though the multilevel models have been shown to be useful, Curran demonstrated that a two-level multilevel model with both level-1 and level-2 predictors and fixed and random effects can be estimated within the standard SEM framework, but the data management problem was remarkably complex, tedious, and error prone. However, this complexity is obviated with modern software (Muthen & Muthen, 2005).

SEM is primarily aimed at studying the relationships among sets of variables, which can be either observed or unobserved. SEM is used as a confirmatory more than an exploratory modeling method, and thus allows researchers to test hypothesized models and modify them subsequently according to theory and sample-based evidence. As a confirmatory technique, SEM requires a substantive theory underlying the hypothesized model and a representative sample for data analysis. When the model fit is not satisfactory, theoretical justifications are needed to revise the model in addition to the mere statistical modification indices (Hancock & Mueller, 2006).

SEM offers several advantages over standard multilevel modeling software programs. Namely, one can model multiple indicator level-2 predictors and multiple indicator dependent variables. According to Curran (2003), measurement models have been proposed in the standard HLM approach, which are currently limited in that pattern coefficient must be fixed to unity and all residual variances are equated for all items. In SEM, no such limitations exist.

Fit Indices in SEM

One current challenge today is the difficulty of calculating an omnibus measure of model fit of multilevel models outside the SEM framework. This difficulty is due to the fact that no logical saturated model is available with which to compare the fitted model. In contrast, the SEM design allows for a natural saturated model (baseline model) to which any fitted model can be compared. Although any model nested under the target model or model of interest may serve as the comparison model, the independence model is most often compared. The primary assumptions of the independence model are the observed variables are measured without error (i.e., all error variances are fixed to zero and pattern coefficients are fixed to one) and all variables are uncorrelated. Note the independence model is restrictive as only the variances of the variables are estimated. Additionally, one may fit a null model to the data. In the null model, all parameters are fixed to zero with no parameters being estimated (Jöreskog & Sörbom, 1993, p. 122). The fit index for a baseline model (independence or null model) will usually indicate a bad fit of the model to the data and serve as a comparison value. The issue is whether the target model or model of interest is an improvement relative to the baseline model.

To determine if the target model is an improvement over the baseline model, one should examine the likelihood ratio test (LRT). According to Curran (2003) and Shumaker and Lomax (1996), the LRT is based on a chi-square distribution that allows for a variety of fit indices which are not available outside the SEM framework. This is one advantage of using SEM to calculate multilevel models. More often than not, the multilevel model calculated outside the SEM framework is assumed to provide an adequate fit when in fact there may be serious flaws with the target model.

Nonetheless, the chi-square statistic is used to evaluate the appropriateness of the structural equation model by determining whether the population covariance matrix is equal to the model-implied covariance matrix. In other words, the LRT tests the hypothesis that the differences between the population covariance matrix and the model-implied covariance matrix are zero. Although the population covariance matrix is rarely known, researchers must examine the sample or empirical covariance matrix and the model-implied covariance matrix. If the null hypothesis is correct, the minimum fit function times N-1 converges to a chi-square variate (Bollen, 1989).

According to Fan, Thompson, and Wang (1999) and Hox (2002), there are several disadvantages of the chi-square statistic that include:

1) Violation of assumptions. The chi-square statistic is based on the assumption that the observed variables are multivariate normal and the sample is sufficiently large. However, these assumptions are rarely met in many practical applications, especially educational data where there is a propensity for the data to exhibit extreme kurtosis and skewness and varying sample size.
2) Model complexity. The chi-square value increases when additional parameters are included in the model. Thus, the chi-square values of more complex models tend to be smaller than the simpler models based on the reduction in the degrees of freedom.
3) Dependence on sample size. As sample size increases while maintaining the same degrees of freedom, the

> chi-square value increases. This could lead to plausible models being rejected based on a statistically significant chi-square test when the discrepancy between the sample and model-implied covariance matrix is in fact trivial.

On the other hand, as sample size decreases, the chi-square value decreases as well, and the LRT may indicate that the discrepancy between the sample and model-implied covariance matrix is near zero, when in fact, the discrepancy is considerable. Based on these shortcomings, Jöreskog and Sörbom (1993) suggested not using the chi-square statistic as a formal test statistic but rather as a descriptive goodness-of-fit index. As an alternative, Jöreskog and Sörbom suggested that one should compare the magnitude of the chi-square value with the expected value of the sample distribution (i.e., the degrees of freedom). For a good model fit, the ratio χ^2/df should be as small as possible with values of 2 or 3 indicative of an acceptable fit. However, according to Bollen (1989) 'the problem of sample size dependency cannot be eliminated by this procedure' (p. 278).

Common Fit Indices

Fortunately, due to the sensitivity of the chi-square statistic to sample size, alternative goodness-of-fit measures have been developed that include the Root Mean

Square Error of Approximation (RMSEA), Root Mean Square Residual (RMR), Standardized Root Mean Square Residual (SRMR), Comparative Fit Index (CFI), Nonnormed Fit Index (NNFI), Goodness of Fit (GFI)and the Adjusted Goodness of fit (AGFI) indexes, which are available in the SEM framework to evaluate the utility of the multilevel model.

The RMSEA is a measure of the approximate fit in the population and is concerned with the discrepancy due to approximation (Steiger, 1990). The RMSEA is bound below by zero. According to Steiger(1990) and Browne and Cudeck (1993), a "close fit" is a RMSEA value less than or equal to .05. Further, Browne and Cudeck consider RMSEA values < .05 a good fit, values between .05 and .08 as an adequate fit, and values between .08 and .10 as a mediocre fit, whereas values > .10 were not acceptable. Although there is general agreement in the field that the value of RMSEA good model fit should be .05 or less, Hu and Bentler (1999) suggested an RMSEA cutoff value of less than .06 as an indication of good fit of the model to the data.

The RMR index is an overall badness-of-fit measure that is based on the fitted residuals (Jöreskog and Sörbom, 1989). In principle, RMR values close to zero suggest a good fit. However, the elements of sample and model-implied covariance matrices are scale dependent, which suggests that the fitted residuals are scale dependent, too, implying that RMR depends on the magnitude of the variances and covariances of the observed variables. In other words, without taking the scales of the variables into account it is virtually impossible to say whether a given RMR value indicates good or bad fit. To overcome this dilemma, Bentler (1995) introduced the SRMR.

As the RMR and the SRMR are overall measures based on squared residuals, they can give no information about the directions of discrepancies between the sample and model-implied matrices. In a residual analysis, it is important to take the sign of a residual into account when looking for the cause of model misfit. Given that an empirical covariance is positive, a positive residual indicates that the model underestimates the sample covariance. In this case, the empirical covariance is larger than the model implied covariance. A negative residual indicates that the model overestimates the sample covariance, that is, the empirical covariance is smaller than the model-implied covariance.

Bentler (1989), Bollen (1989) and Jöreskog and Sörbom (1989) reported that model fit and statistical significance tests may be affected by deviations from normality. Fan, Thompson, and Wang (1999) reported that the RMSEA is sensitive to model misspecification and, along with the Comparative Fit Index (CFI) and Nonnormed Fit Index (NNFI), is minimally influenced by sample size. The CFI ranges from zero to one with higher values indicating better fit. A rule of thumb for the CFI index is that .97 or greater is indicative of good fit relative to the independence model, while values greater than .95 may be interpreted as an acceptable fit. On the other hand, the NNFI, which is also known as the Tucker-Lewis Index (TLI) (Schermelleh-Engel, Moosbrugger & Müller, 2003) generally ranges from zero to one. This index is not normed and values can sometimes occur outside this range, with higher NNFI values indicating better fit. As commonly practiced, an index value of .97 is indicative of good fit relative to the independence model, whereas values greater than .95 may be interpreted as an acceptable fit.

Fan et al. (1999) reported that two commonly applied fit indices, namely the Goodness of Fit Index (GFI) and the Adjusted Goodness of Fit Index (AGFI), are sample size dependent, which was in agreement with earlier studies (cf. Hu & Bentler, 1995, 1998, 1999). The GIF (Jöreskog & Sörbom, 1989; Tanaka & Huba, 1984) measures the relative amount of the variances and covariances in the empirical or sample covariance matrix that is predicted by the model-implied covariance matrix. According to Jöreskog and Sörbom (1993), this implies testing how much better the model fits as compared to ‘no model at all’ (null model) (i.e., when all parameters are fixed to zero) (p. 123). In addition, Jöreskog and Sörbom (1989) also developed the AGFI to adjust for a bias resulting from model complexity. The AGFI adjusts for the model's degrees of freedom relative to the number of observed variables and therefore rewards less complex models with fewer parameters.

Finally, Fan et al. (1999) found that the GFI, AGFI, and the RMSEA were not overly influenced by the estimation method (ML or GLS) as evidenced by the total amount of variation attributed to the estimation method. Although these fit indices were examined in the context of factor analysis, there has been relatively little research examining how these fit indices perform for multilevel regression models conducted in the SEM framework. Further, to the researcher’s knowledge, no studies have been conducted that examine how these fit indices perform under various estimation methods (i.e. ML, GLS, and WLS) when data exhibit departure from normality under various sample sizes in multilevel regression models in the SEM framework.

Akaike (1973) developed the Akaike Information Criterion (AIC), which is an *estimate* of the expected, relative distance between the fitted model and the unknown true mechanism that actually generated the observed data. In application, one computes AIC for each of the candidate models and selects the model with the smallest value of AIC. The model with the smallest AIC value is the model that is estimated to be closest to the unknown reality that generated the data, from among the candidate models considered (Burnham & Anderson, 1998). The importance of theory guiding model construction cannot be underestimated as models not in the set being investigated remain out of consideration. In more succinct terms, Burnham and Anderson reminded researchers that the AIC is useful in selecting the best model in the set; however, if all the models are very poor, AIC will still select the one estimated to be best, but even that model may be poor in an absolute sense. Thus, every effort must be made to ensure that the set of models is well founded.

Regarding interpretation, AIC is usually positive, however, AIC can be shifted by any additive constant, and some shifts can result in negative values of AIC. Computing AIC from regression statistics often results in negative AIC values. Note however, that it is not the absolute size of the AIC value, it is the relative values over the set of models considered, and particularly the differences between AIC values, that are important (Burnham & Anderson, 1998).

As with most statistics, there are disadvantages to AIC. According to Bozdogan (1987), AIC which does not directly involve the sample size, therefore has been criticized as lacking properties of consistency. To counter, Schwarz (1978) developed the Bayesian Information Criterion (BIC) a popular alternative to AIC presented by Schwarz (1978) and Akaike (1978) that does incorporate sample size in its estimates. Studies comparing both the

AIC and BIC found that when the order of the model is known and for reasonable sample sizes, there is a tendency for AIC to select models that are too complex and for BIC to select models that are too simple (Gagne, n.d.; McQuarrie & Tsai, 1998).

SEM is rapidly gaining popularity in the social sciences (Curran, 2003; Tomarken & Baker, 2003). A plausible explanation for the increased use of SEM is that user-friendly software packages have made SEM accessible to substantive researchers who may have limited training in statistics (e.g., Arbuckle & Wothke, 1999; Bentler, 1995; Jöreskog & Sörbom, 1996; Jöreskog, Sörbom, Du Toit, & Du Toit, 2000; Muthén & Muthén, 2001; Steiger, 1995).

Mehta and Neale (2005) reported that multilevel SEM models represent a considerable advance over conventional multilevel models when considering that most applications of multilevel models are restricted to univariate outcomes. Perhaps one of the most distinct advantages of the multilevel model is that there are well developed methods for the incorporation of alternative link functions to allow for explicit modeling or dependent measures that are scaled as dichotomous, ordinal or count variables. However, the estimation of the standard multilevel model under maximum likelihood assumes that any continuous dependent variables are normally distributed (Curran, 2003). This is especially important considering that educational data are often not normally distributed. Indeed, variable skew appears to be most problematic in behavioral research. Micerri (1989) reported that among 440 large-sample achievement and psychometric measures, 71.6% showed moderate to extreme skew for their score distributions while some measures showed even greater likelihood of asymmetry.

Tomarken and Waller (2005) compared multilevel regression modeling and multilevel SEM and noted the advantages of multilevel SEM to be interpretable measures of goodness of fit, better modeling of residuals, and better capacity to model latent variables. Regarding multilevel regression, the researchers reported that the advantages of multilevel regression include easier model specification, fewer estimation problems, and the ability to handle certain types of analysis difficult to handle within multilevel SEM. For instance, in multilevel SEM one cannot model between-group variability in factor pattern or path coefficients (Hox, 2002). Tomarken and Waller noted, however, that both multilevel regression and multilevel SEM are more similar than different. The similarities are further underscored by Bollen (1989) and Hox (2002) who noted that HLM and SEM are closely related in the sense that HLM can be theoretically specified as a special case of the more general structural equation model as a linear model whose covariance structure can be written in finite form. Nonetheless, a recent review of the literature in educational leadership found that the multilevel SEM is becoming the preferred method to examine multilevel data (Byrd, 2007).

The recent upsurge of interest in SEM does, however, have a potential downside. As Steiger (2001) has pointed out, one unfortunate consequence of its increased popularity and accessibility may be lack of awareness on the part of many users of its assumptions, limitations, complexities, and ambiguities. Given the frequency of its use and the automation of many salient components of the SEM model, it is important to recognize the assumptions associated with different estimation methods.

Assumptions and Estimation Methods in SEM

The importance of adhering to the assumptions and selecting appropriate strategies and estimators to model the data cannot be ignored. Violations of the assumptions of SEM (i.e., multivariate normality, sample and cluster size) can lead to biased results that may prompt incorrect decisions about the theory or model being tested (Finney & DiStefano, 2006). As with most statistics, and particularly SEM, there are assumptions that must be met in order for researcher to trust the obtained results. A primary consideration in SEM is the choice of the estimation method that is used to obtain parameter values, standard errors, and fit indices (Brown, 2006). Regarding normal theory, two popular estimators include maximum likelihood (ML) and generalized least squares (GLS) (Muthen & Muthen, 1998-2006). These common estimators require the following assumptions (Bollen, 1989; Kaplan, 2000; Muthen & Muthen, 2004-2006):

1) Independent observations. Observations for different subjects are independent.
2) Large sample size. All statistics estimated in SEM are based on asymptotic properties, meaning that a sufficiently large sample size is required.
3) Correctly specified model. The model should represent the true structure in the population.
4) Multivariate normal data. The observed scores should have a multivariate normal distribution.
5) Continuous data. Although there are those who argue that categorical data can be continuous data as they have equal intervals, this is not the case. According to Bollen (1989) and Kaplan (2000), categorical data by definition cannot be continuous. As the authors submitted, dichotomized data and Likert-type data cannot be normally distributed as these measures are discrete in nature. Therefore, it is often noted that normal theory estimators require continuous normally distributed endogenous variables.

If the SEM assumptions are met, the resulting parameter estimates are asymptotically unbiased (meaning that the estimates neither over-nor underestimate the true population parameters), efficient (minimum variability of the parameter estimates), and consistent (parameter estimates converge to population values as sample size increases).

Defining Normal Theory Estimators

The maximum likelihood estimator is the default in most statistical packages and is used most often. The maximum likelihood fit function is provided in equation 1 below

$$\text{ML} = \frac{1}{2} tr[([S - \Sigma(\theta)]W^{-1})^2)] \qquad [1]$$

where *tr* is the trace of a matrix, which is the sum of the elements on the diagonal, and $S - \Sigma(\theta)$ is the difference between the sample covariance matrix and the model implied matrix. The residuals from $S - \Sigma(\theta)$ are weighted by a weight matrix (**W**). It is the weight matrix that differs between the ML and GLS estimators. In maximum likelihood, the model implied covariance matrix $\Sigma(\theta)$ is the weight matrix, and with GLS the sample covariance matrix **S** is employed as the weight matrix. Browne (1884) has shown that different estimation procedures such as ML, GLS, and the asymptotically distribution free weighted least squares estimator (WLS), which will be discussed shortly, produce estimates that converge and possess the asymptotic properties previously described. In more succinct terms, under ideal conditions, the choice of an estimation method is arbitrary (Ollson, Foss, Troye, & Howell, 2000).

However, the maximum likelihood estimator assumes that the variables in the model are multivariate normal (i.e., the joint distribution of the variables is a multivariate normal distribution). Furthermore it is assumed that $\Sigma(\theta)$ and **S** are positive definite, which means that these matrices must be nonsingular. Maximum likelihood estimators have several important properties (cf. Bollen, 1989). If the observed data stem from a multivariate normal distribution, if the model is specified correctly, and if the sample size is sufficiently large, maximum likelihood provides parameter estimates and standard errors that are asymptotically unbiased, consistent, and efficient. Furthermore, with increasing sample size the distribution of the estimator approximates a normal distribution. Thus, the ratio of each estimated parameter to its standard error is approximately *z*-distributed in large samples.

An important advantage of maximum likelihood is that it allows for a formal statistical test of overall model fit for over identified models. The asymptotic distribution of (ML) F_{ML} is a χ^2 distribution with $df = s - t$ degrees of freedom, where s is the number of nonredundant elements in **S** and t is the number of free parameters. Another advantage of maximum likelihood is that its estimates are in general scale invariant and scale free (Bollen, 1989, p. 109). As a consequence, the values of the fit function do not depend on whether correlation or covariance matrices are analyzed, and whether original or transformed data are used (Schermelleh-Engel, Moosbrugger, & Müller, 2003).

A limitation of maximum likelihood estimation is the strong assumption of multivariate normality, as violations of distributional assumptions are common and often unavoidable in practice and can potentially lead to seriously misleading results. Nevertheless, maximum likelihood seems to be quite robust against the violation of the normality assumption (cf. Boomsma & Hoogland, 2001; Chou & Bentler, 1995; Curran, West & Finch, 1996; Muthén & Muthén, 2002; West, Finch, & Curran, 1995). Simulation studies suggest that under conditions of severe nonnormality, maximum likelihood parameter estimates are still consistent but not necessarily efficient.

Generalized Least Squares (GLS) is a frequently used estimation method that is asymptotically equivalent to ML. As GLS is based on the same assumptions as maximum likelihood, this estimation method is used under the same conditions. However, according to Olsson, Foss, and Howell

(2000), GLS performs poorly in small samples. Therefore, they recommend that the maximum likelihood estimator be utilized with small sample sizes.

Generally, the effects of non-normality on normal theory estimators (ML & GLS) depend on the extent of non-normality, with increasing non-normality leading to a greater impact on the obtained results (Finney & DiStefano, 2006). Consequently, it is imperative that researchers examine the distribution of the observed variables prior to beginning the SEM analysis. Three indices of non-normality are typically employed, univariate kurtosis, univariate skewness, and multivariate kurtosis. Although consensus has not been reached among researchers regarding a clear cut value of non-normality, Chou and Bentler (1995) and West, Finch and Curran (1995) suggested that one may experience problems with maximum likelihood estimators when any univariate skew approaches two and any univariate kurtosis nears seven. Further, regarding multivariate kurtosis, it appears that there is no generally accepted value indicating non-normality. However, in prior research, Bentler and Wu (2002) found that values in excess of three obtained from Mardia's normalized multivariate kurtosis equation (Bollen, 1989, equation 2) could lead to inaccurate results when employing the maximum likelihood estimator. Note in multivariate statistics, the multivariate normality assumption will not hold if any of the analysis variables violate univariate normality (Bollen, 1989).

Muthen and Kaplan (1985) argued any nonnormal conditions with the absolute values of skewness and kurtosis less than 1 would not cause severe distortion in computational results. According to Zhang (2005), transformations can reduce the degree of skewness and kurtosis with the degree of kurtosis much harder to correct. Regarding bias, Chou, Bentler, and Satorra (1991) and Muthen and Kaplan (1985) agreed that data with both nonzero skewness and kurtosis tend to produce more biased results than when data exhibit departure from normality in either skewness or kurtosis using normal theory estimators. In a simulation study, Lei and Lomax (2005) examined the effects of the maximum likelihood and generalized least squares estimators in varying sample size and data normality conditions on standard error bias in structural equation models. In their study, data normality was defined as skewness equals 0 and kurtosis equals 0, slight nonnormality was defined as when skewness equals between 0.3 and 0.4 and kurtosis equals around 1.0. In addition, severe normality was defined as when skewness was above 0.7 and kurtosis was above 3.5. The results revealed that standard errors of parameter estimates were not significantly affected by estimation methods and nonnormality conditions. As expected, standard errors decreased in larger sample sizes. Parameter estimates were more sensitive to nonnormality than to sample size and estimation method.

Because SEM models are commonly applied to nonnormal data, many efforts have been made to evaluate estimation bias in parameters and standard errors as well as efforts to develop estimation procedures and test statistics robust to nonnormality. Studies have examined the influence of nonormailty on parameter estimates, standard errors, and fit indices in factor analysis in the SEM framework using Monte Carlo simulation studies (e.g. Olsson, Foss, & Howell, 2000). Chou, Bentler, and Satorra (1991) concluded that the performance of GLS and ML with respect to empirical fit was reasonably robust to moderate deviations from multivariate normality. But Bentler (1989), Bollen (1989), and Jöreskog and Sörbom (1988) reported that model fit and significance tests may be affected by deviations from normality.

In structural models, as opposed to functional models, all variables are taken to be random rather than having fixed levels. For maximum likelihood (default) and generalized least-squares estimation, the random variables are assumed to have an approximately multivariate normal distribution. Nonnormality, especially high kurtosis, can produce poor estimates and grossly incorrect standard errors and hypothesis tests, even in large samples.

Defining Nonnormal Theory Estimators

When the data do not meet the assumptions for normal theory estimators, namely observed endogenous variables exhibit departure from normality, observations are not independent, or the data are categorical in nature, one may employ weighted least squares (WLS) which is an asymptotic distribution free (ADF) estimator (Browne, 1984). The WLS estimator is commonly defined as

$$\mathrm{WLS} = (\mathbf{S} - \hat{\sigma})' W^{-1} (\mathbf{S} - \hat{\sigma}) \qquad [2]$$

where **S** represents a matrix of non-duplicated elements in the sample covariance matrix(S), $\hat{\sigma}$ is a vector of non-duplicated elements in the model implied covariance matrix, and **S** - $\hat{\sigma}$ is the difference between the sample values and the model-implied values. In similar fashion to the normal theory estimators, the residuals (**S** - $\hat{\sigma}$) are weighted by a weight matrix W. The W matrix is an asymptotic covariance matrix that is calculated using the sample covariance matrix and fourth-order elements or measures of kurtosis (Finney & Distefano, 2006).

The WLS method has advantages as well as disadvantages (Bollen, 1989). One main advantage is that it requires only minimal assumptions about the distribution of the observed variables. Simulation research with nonnormal data shows that the WLS test statistic is relatively unaffected by distributional characteristics (Hoogland & Boomsma, 1998; West, Finch, & Curran, 1995). Another advantage is that WLS may also be used as a means of analyzing correlation matrices, if the corresponding matrix W contains the covariances of the correlations.

In general, WLS produces an accurate χ^2 test statistic and accurate standard errors if sample size is sufficiently large. Boomsma and Hoogland (2001) found that for a sample size of $N \leq 200$, "ADF is a disaster with more than one-third of the solutions being improper" (p. 148). For nonnormally distributed variables, recommendations are quite divergent. Minimum sample size for WLS estimation should be at least 1,000 (Hoogland & Boomsma, 1998), and depending on the model and data analyzed, in some cases even more than 4,000 or 5,000 cases are required (Boomsma & Hoogland, 2001; Hoogland, 1999; Hu, Bentler & Kano, 1992). Jöreskog and Sörbom (1996) recommended that a minimum sample size for employing WLS is 1.5p * (p +1) for p > 12 where p = the number of observed items in a given model. However, results based on small to medium sample sizes should be interpreted with caution (Bollen, 1989).

A disadvantage of the WLS method can be seen in the fact that the weight matrix grows rapidly with increasing numbers of indicator variables. As the asymptotic covariance matrix is of order (k x k), where $k = p(p + 1)/2$ and p is the number of observed variables, the weight matrix of a model containing 10 variables would be of order (55 x 55) with 3025 non-redundant elements. Thus, the WLS method compared to maximum likelihood requires large samples in order to obtain consistent and efficient estimates. If the distribution of the observed variables does not deviate from the normal distribution by a considerable amount, one may also apply maximum likelihood. Consistent with previous findings (cf. Chou, Bentler, & Satorra, 1991; Muthén & Kaplan, 1985, 1992), Chou and Bentler (1995) did not recommend WLS for practical applications when models are complex and when the sample size is small. Further, simulation studies suggest that the maximum likelihood estimator with or without correction for nonnormality seems to perform better than WLS and should be the preferred method (Boomsma & Hoogland, 2001; Hu, Bentler, & Kano, 1992; Olsson, Foss, Troye, & Howell, 2000).

More recently, Olsson, Foss and Howell (2000) studied the effects of three estimation methods (ML, GLS, & WLS) on indexes of fit and parameter bias for different sample sizes when nested models vary in terms of specification error and differing levels of kurtosis. The results revealed that ML, compared to GLS, under conditions of misspecification provides more realistic indexes of overall fit and less bias in parameter estimates for paths that overlapped the true model. Further, they found that WLS was not preferable to either ML or GLS in terms of parameter bias or model fit even though with large samples sizes and mildly misspecified models, WLS provided estimates close to the ML and GLS estimators. However, their study was limited to factor analysis and not multilevel regression models in the SEM framework.

Summary

Chapter II provided a review of the literature that included an overview of the multilevel model, commonly reported fit indices and estimators that are available in the SEM framework. To date, as evidenced in Chapter II, investigation of the impact on estimation methods (WLS, GLS, & WLS) on parameter estimates, standard errors and fit indices have been limited to factor analysis while virtually no studies (except Zhang, 2005) have examined how these methods perform in multilevel models. With the recent advancement of multilevel models in the SEM framework, it is not clear how different estimation methods parameter estimates, standard errors and fit indices perform under varying sample and cluster sizes with normal and nonnormal data. With the advancement of statistical packages automating many of the salient components of SEM modeling (cf. Mplus, AMOS), and given that most data in educational settings are not normally distributed (Micerri, 1989), it is important that researchers understand how different estimators impact key elements of the multilevel model in the SEM framework.

CHAPTER III

METHODOLOGY

The purpose of this Monte Carlo study was to determine how maximum likelihood, generalized least squares and weighted least squares estimators perform in multilevel regression in the SEM framework when data exhibit departures from normality, which is common in educational data (Micerri, 1989). Three two-level models including (1) a model with one level-one predictor and one level-two predictor, (2) a model with two level-one and one level-two predictors, and (3) finally, a model with two level-one and level-two predictors were analyzed. The derivation of these models was intended to represent a real scenario in education where students are nested within classrooms.

Rationale for Simulation Model

The number of predictor variables at each level was based on earlier work by Hox (2000), who examined one predictor at both level-one and level-two, and Bryk and Raudenbush (1992, 2000), and Singer (1998) who examined two level-one variables and two and four level-two predictors in subsequent analysis when examining the *High School and Beyond* national dataset, which has become a classic in HLM research (Gibson, 2003). The present study was not intended to cover all possible data structures, but rather to simulate data that have a structure which is close to the data structures that have appeared in similar studies involving substantive results. Table 1 displays a representative sample of prior research that has investigated the utility of hierarchical linear models and the number of predictor variables at both level-one and level-two.

Table 1
A Representative Sample of Prior HLM Studies Comparing the Number of Predictor Variables at both Level-One and Level-Two

Author	Number of Level-1 Predictors	Number of Level-2 Predictors	Purpose of Study
Zahng(2005)	1	1	Monte Carlo
Gibson (2003)	2	2 & 4	Monte Carlo
Hox (2002)	1	1	Heuristic Example
Raudenbush and Bryk(2000)	2	2 & 4	Heuristic Example
Singer (1998)	2	2 & 4	Heuristic Example
Kreft, de Leeus, & van der Leeden (1994)	Example #1 1 Example #2 2	Example #1 1 Example #2 1	Monte Carlo
Bryk and Raudenbush(1992)	2	2 & 4	Heuristic Example

The conditions investigated included varying degrees of sample size at level one and varying the number of clusters in normal and nonnormal data conditions and the interaction of both sample size and number of clusters and the effect on parameter estimates, standard errors and fit indices. In addition, the series of models investigated in the present study increased in complexity and were designed to determine how adding additional predictors at both levels (coupled with data nonnormality, estimation method, and sample size) impacted bias among parameter estimates, standard errors, and fit indices. The purpose of the study was to provide guidance regarding the use of different estimators in multilevel SEM when researchers encounter similar data.

Data Generation

Initially, normally distributed population correlation matrices with five variables (x1, x2, W1, W2 and y) were generated for each of the following conditions examined. Each variable maintained a mean of zero and a standard deviation of one. The correlation between each variable was initially set at .3. A correlation of .3 was chosen as this is indicative of a moderate effect that is common in the behavioral sciences (Cohen, 1988; Zahng, 2005). However, correlations of .5 and .7 were also examined to determine how the different estimators performed in terms of bias of parameter estimates, standard errors and fit indices.

Based on earlier work by Zahng (2005), generated data was initially sorted by W_n, which is the level-two predictor variable. After being sorted by W_n the scores were divided into classrooms for a given cluster size. The cluster means of W_n were then calculated from the individual values of W_n within each cluster to serve as the level-2 predictor. Two individual predictors (x1 and x2) and two cluster-level predictors, which is the cluster mean of W_n, namely $W_n$bar, together with the cross-level interaction between them (i.e., x1*W_1bar and x2*W_2bar) were utilized to predict the individual level outcome Y. Three different models were investigated to determine how adding additional predictor variables at both level-1 and level-2 and cross-level interactions impact parameter estimates, standard errors and fit indices bias. The models examined in the study are defined below.

Model One

Based on earlier work by Zahng (2005), Hox (2002), and Kreft, de Leeus, and van der Leeden (1994), this model included one predictor variable at each level of the model.

Level-1 equation $Y_{ij} = \beta_0 + \beta_1 X1_{ij} + \varepsilon_{ij}$

Level-2 equation $\beta_0 = \gamma_{00} + U_{0j}$

$\beta_1 = \gamma_{10} + \gamma 11X\ W_1BAR_j$

Combined Equation $Y = \gamma_{00} + \gamma_0 X_1 + \gamma_{11} W_1BAR*X1 + U_0 + \varepsilon$

Model Two

Following the examples of Gibson (2003), Raudenbush and Bryk (2000), Singer (1998) and Bryk and Raudenbush (1992), this model included two predictor variables at level-1 and one predictor variable at level-2.

Level-1 equation $Y_{ij} = \beta_0 + \beta_1 X1_{ij} + \beta_2\ X2_{ij} + \varepsilon_{ij}$

Level-2 equation $\beta_0 = \gamma 00 + U0_j$

$\beta_1 = \gamma 10 + \gamma_{11}\ W_1BAR$

Combined equation $Y = \gamma_{00} + \gamma_0 X1 + \gamma_{11}\ W_1BAR*X1\ U_0 + \varepsilon$

Model Three

Model three is based on the work of Gibson (2003), Raudenbush and Bryk (2000), Singer (1998) and Bryk and Raudenbush (1992), who added two additional predictors to level-2 (for a total of four two-level predictors) while maintaining the initial level-1 predictors (n = 2). Model 3 in the present study included two predictor variables at level-1 and two predictor variables at level-2 and is further defined below.

Level-1 equation $Y_{ij} = \beta_0 + \beta_1 X1_{ij} + \beta_2\ X2_{ij} + \varepsilon_{ij}$

Level-2 equation $\beta_0 = \gamma_{00} + U0_j$

$\beta_1 = \gamma_{10} + \gamma_{11} W_2 BAR$

$\beta_2 = \gamma_{20} + \gamma_{21}\ W_5 BAR$

Combined equation $Y = \gamma_{00} + \gamma_0\ X1 + \gamma_{11}\ W_2 BAR{*}X1 + + \gamma_{20}$ $x2 + \gamma_{21}\ W_5 BAR{*}x2\ U_0 + \varepsilon$

Generating Nonnormal Data

To generate nonnormal data conditions, intermediate pairwise correlations were obtained for each population correlation matrix examined. According to Fan, Felsovalyi, Sivo, and Keen (2003), inter variable correlations and variable nonnormality conditions interact. This interaction causes sample data to deviate from the specified population inter variable correlation pattern. As recommended by Vale and Maurelli (1983), Fleishman’s (1978) power transformation method was utilized to derive the needed intermediate pairwise correlation coefficients.

Initially, Fleishman coefficients for the desired skewness and kurtosis were obtained utilizing the syntax located in Appendix A. Once the coefficients were known, the task of generating nonnormal data was relatively straightforward. The coefficients can be obtained from tables generated by Fleishman or from Fan, Felsovalyi, Sivo, and Keen's (2003) SAS program located in the Appendix.

Equation 3 shows that this method uses a polynomial transformation to transform a normally distributed variable into a variable with specified degrees of skewness and kurtosis.

$$Y = a + bZ + cZ^2 + dZ^3 \qquad [3]$$

Where:

Y is the transformed variable with specified population skewness and kurtosis.

Z is a normally distributed variable with a population mean of zero and a variance of one.

a, b, c, d are coefficients needed for transforming the unit normal variables (Z) to a nonnormal variable with specified degrees of population skewness and kurtosis. Note that a = -c.

Next, the population target correlation coefficient was entered and the intermediate pairwise correlation calculated. After the intermediate pairwise correlations were obtained, they replaced the initial pairwise correlation coefficients. The new coefficients (intermediate coefficients) were then used in calculations to generate multivariate nonnormal Sample data with the desired degree of skewness and kurtosis while maintaining the desired population correlation coefficients.

Design of the Study

The Monte Carlo study included three independent factors that included sample size, data normality conditions, and three estimators for each of the three models described above. Each is discussed below.

Sample Size

Three different level-1 sample sizes were investigated that included 10, 30, and 50 individuals nested in three clusters sizes of 30, 50, and 100. This represented sample sizes of 300, 900, and 1,500, students nested within 30 clusters; 500, 1,500, and 2,500 students nested within 50 clusters; and 1,000, 3,000, and 5,000 students nested within 100 clusters. The sample sizes examined are representative of state and national databases which are readily accessible to empirical researchers.

Data Normality

Data normality conditions were systematically explored for the first level independent variables pooled across all the observations and clusters. Based on earlier work by Lei and Lomax (2005), three levels of data nonnormality were initially considered: normality, slight nonnormality, and severe nonnormality. Normality was defined as when skewness was 0 and kurtosis was 0. Slight nonnormality was defined when skewness was between 0.3 and 0.4 and kurtosis was around 1.0. Severe normality was defined as when skewness was above 0.7 and kurtosis was above 3.5.

In addition to the work by Lei and Lomax, six additional data normality conditions with varying degrees of skewness and kurtosis were investigated. Among the nine total conditions of data normality, skewness ranged from zero to two and kurtosis ranged from zero to seven. The additional data normality conditions were chosen based on earlier studies conducted by Chou and Bentler (1995) and Curran, West and Finch (1995) who suggested that one may experience problems with the ML estimator when univariate skew approaches two and univariate kurtosis nears seven. The descriptive results of the varying data normality conditions investigated are displayed in Table 2.

Table 2
Data Normality Conditions Investigated in the Present Study

Normality Condition	Skewness M	Skewness SD	Kurtosis M	Kurtosis SD
1.00	.000	.051	-.001	.114
2.00	.348	.051	-.002	.133
3.00	.000	.088	.978	.336
4.00	.345	.096	.974	.385
5.00	.749	.050	.001	.138
6.00	.002	.226	3.810	1.404
7.00	1.739	.124	3.987	.949
8.00	-.002	.332	6.517	2.567
9.00	1.960	.225	6.561	2.308

Regarding the descriptive analyses of the parameter estimates, standard errors, and selected fit indices, the data normality conditions were grouped to assist with presentation and interpreting the results. In the descriptive analyses, based on the results from Table 2, condition 1.00 was considered normal, conditions 2.00 through 5.00 were considered moderately nonnormal, and conditions 6.00 through 9.00 were considered severely nonnormal. The means and standard deviations of the three data normality conditions examined in the descriptive analyses are displayed in Table 3. Note that the means of the three data normality conditions examined closely follow the criteria outlined by Lei and Lomax.

Table 3
Data Normality Conditions Investigated in the Descriptive Analyses

Normality Condition	Skewness		Kurtosis	
	M	SD	M	SD
Normal	.000	.051	-.001	.114
Moderately Nonnormal	.360	.275	.488	.559
Severely Nonnormal	.925	.958	5.219	2.333

Estimators

Three different estimators were examined to determine how each performed in varying conditions of sample size, cluster size, and data normality in the multilevel SEM framework. The estimators included maximum likelihood, generalized least squares, and weighted least squares. Although Olsson, Foss and Howell (2000) studied the effects of these three estimation methods on indexes of fit and parameter bias for different samples sizes when nested models vary in terms of specification error and differing levels of kurtosis in factor analysis, little is known how these estimators perform in multilevel regression in the SEM framework.

Dependent Variables

Regarding dependent variables, parameter estimate bias, standard error bias and bias among four common fit indices were compared. Five common fit indices that were examined in the present study included the Goodness-of-fit (GFI), Root mean square residual (RMR), Fit index, Comparative Fit Index (CFI), and the Akaike Information Criterion (AIC). Fan, Thompson, and Wang (1999) found that the CFI are minimally influenced by sample size. Further, they report that the GFI are not overly influenced by the estimation method (ML or GLS). However, there is no clear evidence how these fit indices perform in the multilevel SEM framework with different estimators in varying data conditions.

Data Analysis

Population model parameters, standard errors, and fit indices were obtained from the population correlation matrices and averaged over 10,000 replications. This ensured stability of the population results. Next, 500 replications for each matrix in each of the conditions discussed above was conducted. The data from each replication were saved and averaged to derive the sample parameter estimates, standard errors, and selected fit indices.

Bias for parameter estimates, standard errors, and fit indices were calculated for each replication based on the formula recommended by Brown (2006). While others have used the root mean square of the difference (RMSD) recommended by Gold, Bentler, and Kim (2003) (cf. Zhang, 2005), Brown’s formula was employed to allow the researcher to determine if the detected bias was positive or negative. Bias was calculated in the present study by taking the mean difference of the sample estimates from the population value divided by the population value:

(***population value - sample***)/(population value) [4]

The resulting bias was saved from each replication and averaged over each of the 500 replications to derive the amount of bias for each regression coefficient, standard error, and fit index.

The biases were compared across the independent factors (sample size, data normality, and estimator) for each model investigated with factorial analysis of variance (ANOVA). According to Stevens (2002), factorial ANOVA is especially useful to analyze the effects of more than one independent variable simultaneously on the dependent variable. In the current study, the dependent variable was bias while the independent variables included sample size, cluster size, data normality, and their interactions. Post-hoc comparisons of the statistically significant factors and interactions of the study design factors were conducted. The results of the post-hoc comparison allowed the researcher to determine which design scenario produced results with the least amount of bias.

Summary

The purpose of Chapter III was to discuss the method utilized in this simulation study. In sum, for designing situations involving varying degrees of normality, three estimation methods, and sample sizes, the following steps were taken:

1. Data were generated with varying degrees of normality and population characteristics as discussed above (see Appendix C for SAS syntax utilized to generate data and conduct the simulation analysis).
2. Complete data were analyzed by multilevel SEM with the regression coefficients, standard errors and fit indices output to a data file saved from each of the 500 replications.
3. Regression coefficients, standard errors and selected fit indices from each model were averaged across replications and compared across each of the study’s design effects.
4. A number of separate sets of factorial ANOVAs were conducted to examine the differences in estimation bias among parameter estimates, standard error estimates, selected fit indices.
5. Bias was evaluated by RMSD for parameters, standard errors and selected fit indices.

CHAPTER IV

RESULTS

The purpose of the present study was to compare three estimation methods in multilevel SEM under varying conditions of data normality across different sample sizes to determine which estimator produced more robust parameter estimates, standards, and fit indices. Factorial analysis of variance (ANOVA) was utilized as the primary mode of analysis to evaluate the results of the simulated data. In addition, post hoc analysis was conducted via confidence intervals to provide insight into the statistically significant results. To simplify presentation and interpretation of the results among the six design factors, only the main effects and two-way interactions were reported, even though the full factorial model was studied. Further, only the level-1 main and interaction parameter estimates (level-1 x level-2) few variations under each of the six design factors examined. However, standard errors for each of the parameter estimates (level-1 and level-2) and cross-level interaction effects and their associated standard errors are discussed.

The results of the study are reported in two sections. Section I reports the results of the impact of the six study design conditions on parameter estimate and standard error bias. Section II reports the results of the effect of the six study design conditions on bias associated with the three fit indexes investigated in this study.

Section I: Results of the Impact of the Six Study Design Conditions on Parameter Estimate and Standard Error Bias

Model One

Model one included one predictor variable at level-1 ($\gamma 1$), one predictor variable at level-2, and a cross-level interaction effect (level-1 x level-2). The model is displayed below.

Level-1 equation $Y_{ij} = \beta_0 + \beta_1 X1_{ij} + \varepsilon_{ij}$

Level-2 equation $\beta_0 = \gamma_{00} + U_{0j}$

$\beta_1 = \gamma_{10} + \gamma 11X\ W_1BAR_j$

Combined Equation $Y = \gamma_{00} + \gamma_0 X_1 + \gamma_{11} W_1BAR{*}X1 + U_0 + \varepsilon$

Descriptive Measures of Bias among Parameter Estimates

Tables 4 and 5 display the descriptive results of the level-1 and cross-level interaction parameter estimate bias for model one under three different estimation methods, varying sample sizes, and conditions of data normality. Mean bias among parameter estimates was similar in magnitude across estimation methods for both the level-1 and cross-level interaction term with the greatest amount of bias associated with the cross-level interaction effect. Although greater in magnitude, similar results were reported across estimation methods.

As sample size increased, the amount of bias decreased, which was expected. In addition, as data increased in departure from normality, variability among the bias estimates increased for the interaction effect while variability in bias among the level-1 term remained stable. Note that the results for the level-1 parameter estimate appear to be exact for each estimation method for each condition reported. However, there were slight differences noted when reporting the results to the fourth decimal place.

Table 4

Descriptive Results of Bias among Level-1 Parameter Estimate and Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	.00	.01	.00	.01	.00	.01
	N = 900 (30 x 30)	.00	.01	.00	.01	.00	.01
	N = 1500 (30 x 50)	.00	.00	.00	.00	.00	.00
	N = 500 (50 x 10)	.00	.01	.00	.01	.00	.01
	N = 1500 (50 x 30)	.02	.01	.02	.01	.02	.01
	N = 2500 (50 x 50)	.02	.04	.02	.04	.02	.04
	N = 1000 (100 x 10)	.13	.24	.13	.25	.13	.25
	N = 3000 (100 x 30)	.00	.01	.00	.01	.00	.01
	N = 5000 (100 x 50)	.05	.08	.05	.08	.05	.08
Moderately Nonnormal	N = 300 (30 x 10)	.00	.01	.00	.01	.00	.01
	N = 900 (30 x 30)	.00	.01	.00	.01	.00	.01
	N = 1500 (30 x 50)	.00	.00	.00	.00	.00	.00
	N = 500 (50 x 10)	.00	.01	.00	.01	.00	.01
	N = 1500 (50 x 30)	.02	.03	.02	.03	.02	.03
	N = 2500 (50 x 50)	.03	.06	.03	.06	.03	.06
	N = 1000 (100 x 10)	.13	.26	.13	.26	.13	.26
	N = 3000 (100 x 30)	.00	.01	.00	.01	.00	.01
	N = 5000 (100 x 50)	.05	.10	.05	.10	.05	.10
Severely Nonnormal	N = 300 (30 x 10)	.01	.01	.01	.01	.01	.01
	N = 900 (30 x 30)	.00	.01	.00	.01	.00	.01
	N = 1500 (30 x 50)	.00	.00	.00	.00	.00	.00
	N = 500 (50 x 10)	.00	.01	.00	.01	.00	.01
	N = 1500 (50 x 30)	.00	.02	.00	.02	.00	.02
	N = 2500 (50 x 50)	.00	.05	.00	.05	.00	.05
	N = 1000 (100 x 10)	.01	.34	.01	.34	.01	.34
	N = 3000 (100 x 30)	.00	.01	.00	.01	.00	.01
	N = 5000 (100 x 50)	.04	.08	.04	.08	.04	.08

Table 5
Descriptive Results of Bias among Interaction (L1xL2) Parameter Estimate and Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	-.05	.75	.00	.72	-.01	.71
	N = 900 (30 x 30)	.01	.84	-.02	.95	.03	.78
	N = 1500 (30 x 50)	-.08	.57	-.01	.57	.00	.57
	N = 500 (50 x 10)	-.01	.36	-.01	.35	.00	.36
	N = 1500 (50 x 30)	-.04	.42	-.04	.42	-.03	.45
	N = 2500 (50 x 50)	-.03	.70	.02	.71	-.04	.71
	N = 1000 (100 x 10)	-.02	.42	.00	.42	.00	.42
	N = 3000 (100 x 30)	.01	.44	.00	.47	.02	.48
	N = 5000 (100 x 50)	.05	.79	.03	.81	-.03	.87
Moderately Nonnormal	N = 300 (30 x 10)	.00	.91	.01	.89	.02	.91
	N = 900 (30 x 30)	-.02	2.91	.00	2.90	.01	2.80
	N = 1500 (30 x 50)	-.03	.57	-.04	.58	-.03	.58
	N = 500 (50 x 10)	.03	.45	.02	.43	.02	.45
	N = 1500 (50 x 30)	-.09	.46	-.10	.46	-.08	.46
	N = 2500 (50 x 50)	.03	.71	.05	.71	.04	.72
	N = 1000 (100 x 10)	-.07	.45	-.06	.45	-.06	.45
	N = 3000 (100 x 30)	.01	.53	-.01	.55	.01	.54
	N = 5000 (100 x 50)	-.26	1.05	-.23	1.02	-.22	1.08
Severely Nonnormal	N = 300 (30 x 10)	-.06	1.01	-.03	1.02	-.03	1.01
	N = 900 (30 x 30)	.48	3.09	.56	3.33	.56	3.19
	N = 1500 (30 x 50)	-.09	.99	-.10	.99	-.08	.98
	N = 500 (50 x 10)	-.22	.66	-.22	.65	-.22	.66
	N = 1500 (50 x 30)	-.31	.97	-.32	.99	-.33	.98
	N = 2500 (50 x 50)	-.18	1.20	-.18	1.19	-.17	1.18
	N = 1000 (100 x 10)	.19	.81	.20	.80	.20	.80
	N = 3000 (100 x 30)	-.01	.68	-.02	.68	-.02	.66
	N = 5000 (100 x 50)	-.16	1.27	-.19	1.29	-.16	1.28

Descriptive Measures of Bias among Standard Errors

The descriptive measures of standard error bias for each of the parameters examined are reported in Tables 6 through 8. The results were similar in magnitude across estimation methods with the maximum likelihood and generalized least squares estimators producing less biased standard errors in normal data conditions, while the weighted least squares estimator tended to produce more biased standard errors. Further, regarding the level-1 standard errors in small sample sizes, the standard errors were negatively biased across estimators when the number of clusters ranged from 30-50. However, as the number of clusters increased to 100, bias among the standard errors was positive. In addition, as data normality went from normal to severely nonnormal, the weighted least squares estimation produced standard errors with less bias in small to moderate samples. As the number of clusters increased to 100, all three estimators produced results similar in magnitude regardless of data normality conditions, with the maximum likelihood and generalized least squares estimators producing slightly less biased standard errors than the weighted least squares estimator. Similar results were found among standard errors associated with the level-1, level-2 and the cross-level interaction standard errors respectively

Table 6
Descriptive Results of Bias among Level-1 Standard Error and Selected Design Conditions

Normality	Sample Size (No. Clusters x Cluster Size)	Estimation Method: ML M	ML SD	GLS M	GLS SD	WLS M	WLS SD
Normal	N = 300 (30 x 10)	-1.10	1.61	-1.11	1.63	-1.36	1.92
	N = 900 (30 x 30)	-1.76	2.23	-1.77	2.25	-1.90	2.38
	N = 1500 (30 x 50)	.16	.23	.16	.24	.15	.24
	N = 500 (50 x 10)	-.78	1.87	-.78	1.90	-.98	2.21
	N = 1500 (50 x 30)	-.47	.92	-.47	.93	-.50	.96
	N = 2500 (50 x 50)	-.10	.11	-.10	.11	-.11	.11
	N = 1000 (100 x 10)	.16	.30	.15	.30	.12	.35
	N = 3000 (100 x 30)	.13	.32	.13	.32	.12	.33
	N = 5000 (100 x 50)	.01	.22	.00	.22	-.01	.23
Moderately Nonnormal	N = 300 (30 x 10)	-.35	1.00	-.36	1.04	-.44	1.11
	N = 900 (30 x 30)	-.08	1.04	-.08	1.04	-.08	1.02
	N = 1500 (30 x 50)	-.08	.68	-.08	.68	-.05	.63
	N = 500 (50 x 10)	-.17	.79	-.17	.79	-.24	.89
	N = 1500 (50 x 30)	-.31	1.14	-.31	1.14	-.30	1.11
	N = 2500 (50 x 50)	-.30	.40	-.30	.40	-.27	.37
	N = 1000 (100 x 10)	.17	.23	.17	.23	.15	.26
	N = 3000 (100 x 30)	.14	.29	.14	.29	.14	.30
	N = 5000 (100 x 50)	.13	.22	.13	.22	.13	.23
Severely Nonnormal	N = 300 (30 x 10)	-1.64	7.53	-1.64	7.48	-1.72	7.50
	N = 900 (30 x 30)	-.17	1.65	-.17	1.63	-.06	1.41
	N = 1500 (30 x 50)	.12	.35	.12	.35	.25	.25
	N = 500 (50 x 10)	-.46	1.64	-.47	1.69	-.50	1.85
	N = 1500 (50 x 30)	.01	.26	.01	.26	.07	.21
	N = 2500 (50 x 50)	-.23	.30	-.23	.30	-.11	.19
	N = 1000 (100 x 10)	-.02	.57	-.02	.57	-.04	.59
	N = 3000 (100 x 30)	.12	.26	.12	.26	.20	.25
	N = 5000 (100 x 50)	.20	.12	.20	.12	.28	.06

Table 7
Descriptive Results of Bias among Level-2 Standard Error and Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	-1.19	1.68	-1.21	1.71	-1.09	1.59
	N = 900 (30 x 30)	-1.79	2.25	-1.80	2.28	-1.40	1.80
	N = 1500 (30 x 50)	.16	.24	.16	.24	.17	.21
	N = 500 (50 x 10)	-.88	1.98	-.87	2.00	-.33	.94
	N = 1500 (50 x 30)	-.49	.93	-.49	.94	-.45	.88
	N = 2500 (50 x 50)	-.11	.11	-.11	.11	-.11	.11
	N = 1000 (100 x 10)	.11	.32	.11	.32	.11	.33
	N = 3000 (100 x 30)	.11	.33	.11	.33	.11	.33
	N = 5000 (100 x 50)	.00	.22	.00	.22	-.01	.23
Moderately Nonnormal	N = 300 (30 x 10)	-.41	1.05	-.42	1.09	-.38	1.00
	N = 900 (30 x 30)	-.09	1.05	-.08	1.04	-.06	1.02
	N = 1500 (30 x 50)	-.09	.68	-.09	.68	-.04	.56
	N = 500 (50 x 10)	-.23	.83	-.23	.84	-.25	.89
	N = 1500 (50 x 30)	-.33	1.15	-.33	1.15	-.27	.88
	N = 2500 (50 x 50)	-.31	.40	-.31	.40	-.32	.39
	N = 1000 (100 x 10)	.13	.24	.13	.24	.12	.25
	N = 3000 (100 x 30)	.12	.30	.12	.30	.12	.30
	N = 5000 (100 x 50)	.12	.22	.12	.22	.11	.23
Severely Nonnormal	N = 300 (30 x 10)	-1.75	1.79	-1.73	1.83	-1.51	1.38
	N = 900 (30 x 30)	-.16	1.62	-.16	1.60	-.07	1.09
	N = 1500 (30 x 50)	.13	.34	.13	.34	.12	.34
	N = 500 (50 x 10)	-.52	1.70	-.53	1.76	-.53	1.39
	N = 1500 (50 x 30)	.00	.27	.00	.27	-.03	.29
	N = 2500 (50 x 50)	-.23	.29	-.23	.29	-.26	.31
	N = 1000 (100 x 10)	-.07	.60	-.07	.60	-.10	.62
	N = 3000 (100 x 30)	.10	.26	.10	.27	.09	.27
	N = 5000 (100 x 50)	.19	.13	.19	.13	.18	.13

Table 8
Descriptive Results of Bias among Interaction Standard Error and Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	-1.08	1.60	-1.10	1.62	-1.24	1.79
	N = 900 (30 x 30)	-1.74	2.22	-1.76	2.24	-1.44	1.86
	N = 1500 (30 x 50)	.17	.23	.16	.23	.19	.21
	N = 500 (50 x 10)	-.78	1.88	-.78	1.90	-.38	1.06
	N = 1500 (50 x 30)	-.46	.92	-.46	.92	-.46	.89
	N = 2500 (50 x 50)	-.10	.11	-.10	.11	-.12	.12
	N = 1000 (100 x 10)	.16	.30	.16	.30	.11	.35
	N = 3000 (100 x 30)	.13	.32	.13	.32	.11	.33
	N = 5000 (100 x 50)	.01	.22	.01	.22	-.01	.23
Moderately Nonormal	N = 300 (30 x 10)	-.34	1.00	-.35	1.04	-.40	1.09
	N = 900 (30 x 30)	-.07	1.03	-.07	1.03	-.01	.98
	N = 1500 (30 x 50)	-.07	.67	-.08	.67	.03	.48
	N = 500 (50 x 10)	-.17	.79	-.17	.79	-.25	.97
	N = 1500 (50 x 30)	-.31	1.13	-.31	1.14	-.22	.82
	N = 2500 (50 x 50)	-.29	.40	-.30	.40	-.27	.36
	N = 1000 (100 x 10)	.17	.23	.17	.23	.14	.27
	N = 3000 (100 x 30)	.14	.29	.14	.29	.14	.30
	N = 5000 (100 x 50)	.13	.22	.13	.22	.13	.23
Severely Nonormal	N = 300 (30 x 10)	-1.62	1.69	-1.60	1.44	-1.55	1.87
	N = 900 (30 x 30)	-.14	1.59	-.14	1.58	.06	.94
	N = 1500 (30 x 50)	.14	.34	.14	.34	.24	.26
	N = 500 (50 x 10)	-.35	1.66	-.37	1.71	-.37	1.45
	N = 1500 (50 x 30)	.02	.27	.02	.27	.06	.23
	N = 2500 (50 x 50)	-.22	.29	-.22	.29	-.13	.22
	N = 1000 (100 x 10)	-.02	.57	-.02	.57	-.07	.63
	N = 3000 (100 x 30)	.12	.26	.12	.26	.17	.26
	N = 5000 (100 x 50)	.20	.13	.20	.13	.26	.08

Bivariate Results

Pearson product-moment correlations were calculated to determine how bias among parameter estimates and standard errors was impacted by the number of clusters, cluster size, correlation among variables, and data normality conditions. Regarding parameter estimates, the results reported in Table 9 revealed that the number of clusters was positively correlated with bias in the level-1 parameter estimate (r = .161, P < .01). The 95% CI ranged from 0.156 to 0.165. Yet the number of clusters only accounted for approximately three percent of the variance in bias among the level-1 parameter estimate (r^2 =.025).

Other statistically significant relations among the parameter estimates and study conditions ranged from r = -.085, p < .01 (95% CI ranged from -.089 to -.080) between the level-1 parameter estimate and kurtosis to r =.032, p < .01 (95% CI ranged from .027 to .036) between the cross-level interaction effect and kurtosis. However, the correlations were negligible, accounting for less than one percent of the variance among parameter estimate bias.

Concerning bias among standard errors, the largest correlation was associated with the number of clusters and the standard errors for each of the model-one standard errors. Statistically significant correlations ranged from *r* = -.171, p < .001 between the level-1 parameter estimate standard error and number of clusters (95% CI = -.175 to -.166) to *r* = -.149, p < .001 between the cross-level interaction standard error and the number of clusters (95% CI = -.153 to -.144). To interpret, as sample size increased, bias among each of the parameter estimates decreased. The overall effect size was slight, with the number of clusters explaining approximately three percent of the variation in bias among the standard errors associated with each predictor.

The remaining statistically significant correlations among bias associated with the standard errors ranged from *r* = -.019, p <.01) between the cross-level interaction term and correlation among variables (95% CI ranged from -.024 to -.014) to *r* = -.076, p < 01, between the level-standard error and cluster size (95% CI ranged from -.081 to -.071).

Table 9
Pearson Product-Moment Correlations among Parameter Estimate and Standard Error Bias and Selected Design Conditions

	1	2	3	4	5	6	7	8	9	10	11
Bias Level-1 Predictor (1)	1.00										
Bias Level-2 Predictor (2)	.002	1.00									
Bias Cross-Level Interact.(3)	-.045**	.010**	1.00								
Bias Level-1 Predictor SE (4)	-.030**	.000	-.021**	1.00							
Bias Level-2 Predictor SE (5)	-.035**	-.002	-.011**	.976**	1.00						
Bias Cross-Level Interact. SE (6)	-.038**	-.007**	-.001	.975**	.980**	1.00					
No. of Clusters (7)	.161**	.002	-.020**	-.171**	-.155**	-.149**	1.00				
Cluster Size (8)	-.026**	-.003	-.028**	-.057**	-.076**	-.056**	-.003	1.00			
Correlation among Variables (9)	-.076**	.000	-.004	-.024**	-.019**	-.019**	.000	.000	1.00		
Skew (10)	-.060**	.002	-.021**	.069**	.019**	.036**	-.002	-.002	.000	1.00	
Kurtosis (11)	-.085**	-.002	.032**	.036**	.013**	.034**	-.002	-.002	.000	.403**	1.00

**. Correlation is significant at the 0.01 level (2-tailed). The sample size for each coefficient was 181,875.

Factorial Analysis Results of Parameter Estimates

Table 10 displays the outcome of the factorial ANOVA examining bias in the level-1 parameter estimate. The results revealed that among the main effects, the estimation method was not a statistically significant factor. However, each of the remaining conditions as main effects was statistically significant with the number of clusters accounting for the greatest amount of bias variation in the level-1 parameter estimate (partial η^2 = .019), while the remaining statistically significant main effects explained less than one percent of the total variance in the outcome variable. When examining the two-way interaction effects, the term including cluster size and correlation among variables explained the largest amount of bias variance (approximately six percent) as indicated by partial eta square (partial η^2 = .058). When the estimation method was entered as an interaction term with the remaining design conditions, the results were not statistically significant. Although the remaining interaction effects were statistically significant, each explained less than one percent of the variance in the criterion variable. To provide insight into the results, 95% confidence intervals were calculated. The results are reported in Figure 1.

Table 10
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Level-1 Parameter Estimate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	3E+007	88	303753	485.6	< .001	.190[a]
Intercept	2415354	1	2415354	3862	< .001	.021
Est. Mthd	.225	2	.113	.001	.996	.000
Clstrs	2257604	2	1128802	1805	< .001	.019
Clstr. Size	825690	2	412845	660.0	< .001	.007
Corr	975735	2	487867	780.0	< .001	.009
Skew	286516	3	95505.4	152.7	< .001	.003
Kurt	346184	3	115395	184.5	< .001	.003
Est. Mthd. * Clstrs	.514	4	.128	.002	.975	.000
Est. Mthd. * Clstr. Size	1.214	4	.303	.004	.954	.000
Est. Mthd. * Corr	.749	4	.187	.003	.879	.000
Est. Mthd. * Skew	10.418	6	1.736	.003	.985	.000
Est. Mthd. * Kurt	3.857	6	.643	.001	.958	.000
Clstrs * Clstr. Size	2955847	4	738962	1181	< .001	.025
Clstrs * Corr	2711518	4	677879	1084	< .001	.023
Clstrs * Skew	285174	6	47528.9	75.99	< .001	.003
Clstrs * Kurt	583264	6	97210.6	155.4	< .001	.005
Clstr. Size * Corr	6982636	4	1745659	2791	< .001	.058
Clstr. Size * Skew	161903	6	26983.8	43.14	< .001	.001
Clstr. Size * Kurt	686375	6	114396	182.9	< .001	.006
Corr * Skew	324787	6	54131.2	86.54	< .001	.003
Corr * Kurt	788096	6	131349	210.0	< .001	.007
Skew * Kurt	1784595	2	892297	1427	< .001	.015
Error	1E+008	181786	625.484			
Total	1E+008	181875				
Corrected Total	1E+008	181874				

a. R Squared = .190 (Adjusted R Squared = .190)

Note. The model eta squared for the full factorial 6-way model was .743 with df= 696. The model eta squared for only the main and the two-way interaction effects was .190 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .553 (i.e., .743-.190).

The results displayed in Figure 1 illustrate that as the correlation among variables increased from .30 to .70, variation in bias also increased, especially when cluster size increased from 10 to 50. Further, based on Cummings (2007)“rules of eye”, the overlapping bars for correlation of .30 and .50, respectively, are not statistically significantly different, while the bias associated with the correlation of .70 is statistically significantly different from .30 or .50 when cluster size is held at 30. However, the results should be interpreted with caution as the overall effect size was less than six percent, indicating that the statistically significant results could possibly be attributed to the large sample size.

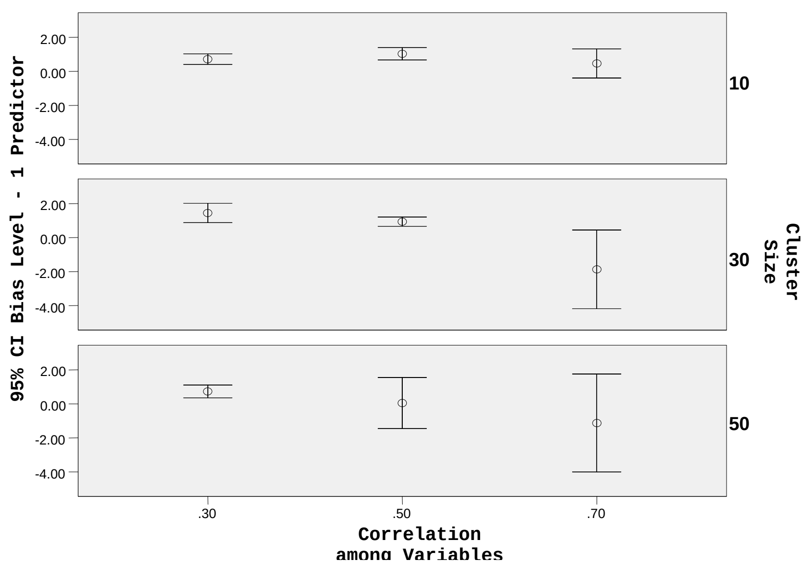

Figure 1. Confidence intervals examining bias among the level-1 parameter estimate in relation to correlation among variables and cluster size.

Table 11 shows that of the six conditions investigated, none had a statistically significant association with bias associated with the level-two parameter estimate either as a main or interaction effect. The results were expected after viewing the results of both the univariate and bivariate analyses reported earlier. Further, according to Mass and Hox (2004), prior studies have shown that the regression coefficients are estimated without bias, while their standard errors tend to be biased downward with small sample sizes, especially at the group level (cf Brown & Draper, 2000).

The results reported in Table 12 regarding the cross-level interaction term are similar to those reported for the level-1 parameter estimate (i.e., the estimation method was not a statistically significant main effect). However, the remaining five conditions did share a statistically significant association with the outcome variable, but explained less than one percent of variance in bias in the cross-level interaction parameter estimate. Similarly, the estimation method, when entered in tandem with the remaining five conditions as an interaction effect, was not statistically significant, while the remaining interaction terms were statistically significant. However, each of the terms explained a negligible portion of variance (less than one percent) in the criterion variable.

Table 11
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Level-2 Parameter Estimate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1.2E+008	88	1E+006	1.654	< .001	.001[a]
Intercept	134454.1	1	134454	.160	.689	.000
Est. Mthd	749787.2	2	374894	.446	.640	.000
Clstrs	1101947	2	550973	.656	.519	.000
Clstr. Size	2136936	2	1E+006	1.272	.280	.000
Corr	755861.7	2	377931	.450	.638	.000
Skew	2606920	3	868973	1.035	.376	.000
Kurt	5086494	3	2E+006	2.019	.109	.000
Est. Mthd. * Clstrs	3716634	4	929158	1.106	.351	.000
Est. Mthd. * Clstr. Size	3249009	4	812252	.967	.424	.000
Est. Mthd. * Corr	4389621	4	1E+006	1.307	.265	.000
Est. Mthd. * Skew	5280552	6	880092	1.048	.392	.000
Est. Mthd. * Kurt	10268618	6	2E+006	2.038	.057	.000
Clstrs * Clstr. Size	3818947	4	954737	1.137	.337	.000
Clstrs * Corr	11395903	4	3E+006	3.393	.009	.000
Clstrs * Skew	3039279	6	506546	.603	.728	.000
Clstrs * Kurt	13057575	6	2E+006	2.592	.016	.000
Clstr. Size * Corr	4080234	4	1E+006	1.215	.302	.000
Clstr. Size * Skew	1606709	6	267785	.319	.927	.000
Clstr. Size * Kurt	7958972	6	1E+006	1.580	.148	.000
Corr * Skew	18990555	6	3E+006	3.769	.001	.000
Corr * Kurt	13320049	6	2E+006	2.644	.015	.000
Skew * Kurt	464034.5	2	232017	.276	.759	.000
Error	1.5E+011	181786	839732			
Total	1.5E+011	181875				
Corrected Total	1.5E+011	181874				

a. R Squared = .001 (Adjusted R Squared = .000)

Note. The model eta squared for the full factorial 6-way model was .007 with df=696. The model eta squared for only the main and the two-way interaction effects was .001 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .006(i.e.,.007-.001).

Table 12
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Cross-Level Interaction Parameter Estimate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	19683.170	88	223.67	155.9	< .001	.070 [a]
Intercept	325.283	1	325.28	226.7	< .001	.001
Est. Mthd	1.918	2	.959	.668	.513	.000
Clstrs	796.637	2	398.32	277.6	< .001	.003
Clstr. Size	473.920	2	236.96	165.1	< .001	.002
Corr	220.866	2	110.43	76.97	< .001	.001
Skew	238.258	3	79.419	55.35	< .001	.001
Kurt	2013.604	3	671.20	467.8	< .001	.008
Est. Mthd. * Clstrs	5.809	4	1.452	1.012	.399	.000
Est. Mthd. * Clstr. Size	.609	4	.152	.106	.980	.000
Est. Mthd. * Corr	2.801	4	.700	.488	.745	.000
Est. Mthd. * Skew	3.233	6	.539	.376	.895	.000
Est. Mthd. * Kurt	2.711	6	.452	.315	.930	.000
Clstrs * Clstr. Size	1386.224	4	346.56	241.5	< .001	.005
Clstrs * Corr	561.581	4	140.40	97.85	< .001	.002
Clstrs * Skew	2362.027	6	393.67	274.4	< .001	.009
Clstrs * Kurt	1199.256	6	199.88	139.3	< .001	.005
Clstr. Size * Corr	959.307	4	239.83	167.1	< .001	.004
Clstr. Size * Skew	821.215	6	136.87	95.39	< .001	.003
Clstr. Size * Kurt	1630.406	6	271.73	189.4	< .001	.006
Corr * Skew	919.784	6	153.30	106.8	< .001	.004
Corr * Kurt	1833.897	6	305.65	213.0	< .001	.007
Skew * Kurt	2028.263	2	1014.1	706.8	< .001	.008
Error	260830.00	181786	1.435			
Total	280712.77	181875				
Corrected Total	280513.17	181874				

a. R Squared = .070 (Adjusted R Squared = .070)

Note. The model eta squared for the full factorial 6-way model was .265 with df= 696. The model eta squared for only the main and the two-way interaction effects was .070 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .195(i.e., .265-.070).

Factorial ANOVA Results of Standard Errors

Table 13 displays the results of the study's design conditions on the standard errors of the level-1 parameter estimates. Regarding main effects, the partial eta squares indicated that the number of clusters, cluster size, and correlation among variables explained the greatest amount of variation in bias. In addition, data conditions regarding skewness and kurtosis were statistically significant; however, each explained little variation in the outcome variable. In regards to the estimation method as a main effect, the estimation method was statistically significant, but explained little variation in the outcome variable.

Interaction effects were examined to gain further understanding of how the conditions examined impacted the dependent variable. The interaction term including the number of clusters and cluster size was statistically significant while having the largest impact on bias among the level-1 standard errors (partial η^2 =.064). In addition, the term including the number of clusters and correlation among variables explained approximately four percent of the variance in the dependent variable (partial η^2 =.043).

Table 13

Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Level-1 Standard Error

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	.138	88	.002	848.2	< .001	.291 [a]
Est. Mthd	.041	1	.041	22073	< .001	.108
Clstrs	.001	2	.000	151.4	< .001	.002
Clstr. Size	.033	2	.017	8951	< .001	.090
Corr	.009	2	.004	2326	< .001	.025
Skew	.009	2	.005	2426	< .001	.026
Kurt	.003	3	.001	509.1	< .001	.008
Est. Mthd. * Clstrs	.004	3	.001	691.5	< .001	.011
Est. Mthd. * Clstr. Size	3.93E-005	4	1E-005	5.292	< .001	.000
Est. Mthd. * Corr	.001	4	.000	121.3	< .001	.003
Est. Mthd. * Skew	1.19E-005	4	3E-006	1.608	< .001	.000
Est. Mthd. * Kurt	.000	6	4E-005	23.66	< .001	.001
Clstrs * Clstr. Size	.000	6	4E-005	22.66	< .001	.001
Clstrs * Corr	.023	4	.006	3108	< .001	.064
Clstrs * Skew	.015	4	.004	2060	< .001	.043
Clstrs * Kurt	.005	6 6	.000	180.3	< .001	.006
Clstr. Size * Corr	.005	4	.001	690.5	< .001	.015
Clstr. Size * Skew	.008	6	.001	692.1	< .001	.022
Clstr. Size * Kurt	.007	6	.001	618.2	< .001	.020
Corr * Skew	.003	6	.001	309.6	< .001	.010
Corr * Kurt	.003	6	.001	298.2	< .001	.010
Skew * Kurt	.002	2	.001	520.5	< .001	.006
Error	.337	181786	2E-006			
Total	.516	181875				
Corrected Total	.476	181874				

a. R Squared = .291 (Adjusted R Squared = .291)

Note. The model eta squared for the full factorial 6-way model was .624 with df=696. The model eta squared for only the main and the two-way interaction effects was .291 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .333(i.e., .624-.291).

Confidence intervals were utilized as a post hoc measure to interpret the salient findings from Table 13. The results displayed in Figure 2 show that as the number of clusters increased, the amount of bias decreased. Note the horizontal line positioned at 0.00 is included to provide a marker indicating no bias. Further, as correlation among variables increased from *r* = .3 to *r* = .7, the standard errors were negatively biased in small sample sizes, especially when the number of clusters was 30 and cluster size was held at 10. As sample size increased, especially when the number of clusters was 100 and cluster size ranged from 30 to 50, the estimation methods produced similar results regardless of data normality conditions.

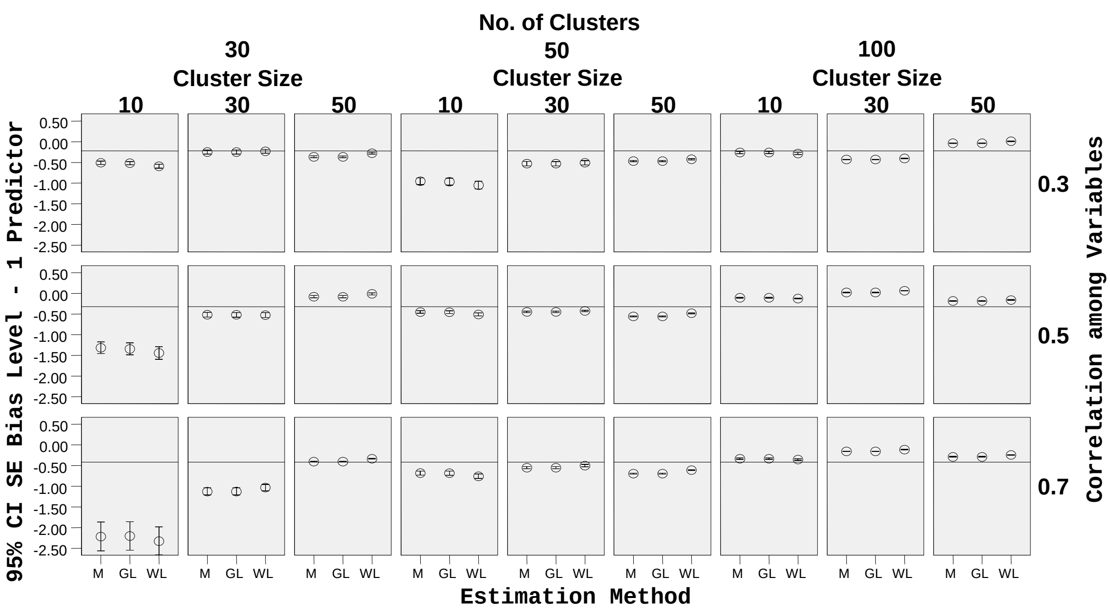

Figure 2. Confidence intervals examining bias among the level-1 standard error by estimation method, number of clusters, cluster size, and correlation among variables.

Table 14 reports the impact of the conditions of the study design on the level-2 parameter estimate's standard error bias. Similar to the level-1 parameter estimate, the number of clusters, cluster size, and correlation among variables appeared to have the largest impact on standard error bias. The number of clusters explained approximately 10%, and cluster size and correlation among variables explained approximately three percent of the variance in the criterion variable. In addition, the estimation method was statistically significant, but explained little variation in the outcome variable. Regarding data normality conditions, skewness and kurtosis were both statistically significant, but explained less than one percent of bias variation based on partial eta squared. When examining interaction effects, the estimation method by the number of clusters was not statistically significant. Statistically significant interaction effects included cluster size by skew, cluster size by kurtosis, and cluster size by correlation among variables. Of these interaction effects, the number of clusters by cluster size explained approximately six percent of the variation in the dependent variable (partial $\eta^2 = .061$).

Table 14
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Level-2 Standard Error

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	.126	88	.001	812.9	< .001	.282[a]
Intercept	.022	1	.022	12496	< .001	.064
Est. Mthd	4E-005	2	2E-005	10.05	< .001	.000
Clstrs	.032	2	.016	8965	< .001	.090
Clstr. Size	.010	2	.005	2925	< .001	.031
Corr	.009	2	.004	2501	< .001	.027
Skew	.002	3	.001	349.1	< .001	.006
Kurt	.002	3	.001	464.6	< .001	.008
Est. Mthd. * Clstrs	1E-005	4	2E-006	1.404	.230	.000
Est. Mthd. * Clstr. Size	3E-005	4	7E-006	3.982	.003	.000
Est. Mthd. * Corr	7E-006	4	2E-006	.986	.414	.000
Est. Mthd. * Skew	6E-005	6	9E-006	5.396	< .001	.000
Est. Mthd. * Kurt	3E-005	6	5E-006	2.790	.010	.000
Clstrs * Clstr. Size	.021	4	.005	2940	< .001	.061
Clstrs * Corr	.015	4	.004	2082	< .001	.044
Clstrs * Skew	.005	6	.001	480.5	< .001	.016
Clstrs * Kurt	.002	6	.000	222.6	< .001	.007
Clstr. Size * Corr	.005	4	.001	679.8	< .001	.015
Clstr. Size * Skew	.008	6	.001	716.2	< .001	.023
Clstr. Size * Kurt	.006	6	.001	597.9	< .001	.019
Corr * Skew	.003	6	.001	317.1	< .001	.010
Corr * Kurt	.004	6	.001	350.2	< .001	.011
Skew * Kurt	.001	2	.000	264.9	< .001	.003
Error	.320	181786	2E-006			
Total	.468	181875				
Corrected Total	.446	181874				

a. R Squared = .282 (Adjusted R Squared = .282)

Note. The model eta squared for the full factorial 6-way model was .618 with df= 696. The model eta squared for only the main and the two-way interaction effects was .282 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .336(i.e., .618-.282).

Figure 3 provides confidence intervals regarding bias in the standard error of the level-2 parameter estimate by estimation method, number of clusters, cluster size, correlation among variables, and kurtosis. The results indicated that as the number of clusters increased, the amount of bias in the dependent variable decreased. This is evident when the correlation among variables ranged from r = .50 to r = .70, and kurtosis increased from four to seven in 30 clusters with a cluster size of 10. Further, the standard errors were negatively biased in small sample sizes (number of clusters = 30 and cluster size ranges from 10 to 30), with the weighted least squares estimator producing slightly less biased results. As sample size increased (namely the number of clusters), the estimation methods produced similar results.

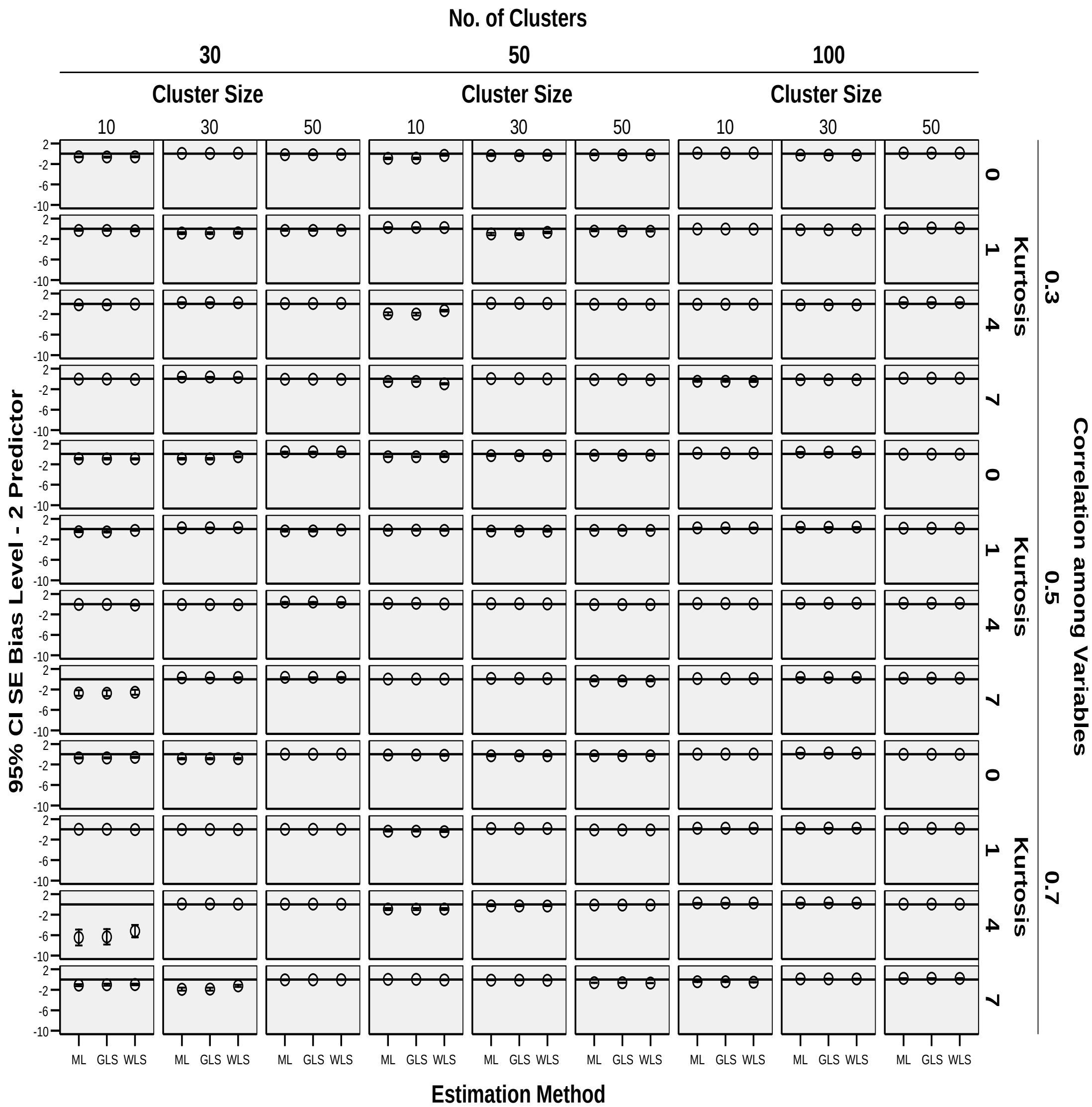

Figure 3. Confidence intervals examining bias among the level-2 standard error, number of clusters, cluster size, correlation among variables, and kurtosis.

The results of the study's conditions on the standard error bias in the cross-level interaction term (level-1 by level-2) are displayed in Table 15. Each main effect was statistically significant with the number of clusters and cluster size explaining the greatest amount of variation in standard error bias. Although the estimation method was statistically significant, the resulting partial eta square was only .002, indicating that the estimation method explained less than one percent of the variation in the outcome variable. Regarding interaction terms, each was statistically significant. Note that the number of clusters by cluster size appeared to have the most noteworthy impact on standard error bias among the cross-level interaction standard error (partial $\eta^2 = .064$). As for data normality conditions, the interaction terms that included cluster size by kurtosis and cluster size by skew shared a statistically significant relation with the dependent variable. These results suggest that clusters with smaller samples in higher degrees of data nonnormality (increased skew and kurtosis) produced negatively biased cross-level interaction standard errors.

Table 15
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Cross-Level Interaction Standard Error

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	.139	88	.002	819.1	< .001	.284[a]
Intercept	.045	1	.045	23646	< .001	.115
Est. Mthd	.001	2	.000	159.2	< .001	.002
Clstrs	.034	2	.017	8809	< .001	.088
Clstr. Size	.010	2	.005	2664	< .001	.028
Corr	.009	2	.004	2268	< .001	.024
Skew	.002	3	.001	383.0	< .001	.006
Kurt	.004	3	.001	634.6	< .001	.010
Est. Mthd. * Clstrs	.000	4	3E-005	16.69	< .001	.000
Est. Mthd. * Clstr. Size	.001	4	.000	153.6	< .001	.003
Est. Mthd. * Corr	2.34E-006	4	6E-007	.305	< .001	.000
Est. Mthd. * Skew	.000	6	4E-005	23.36	< .001	.001
Est. Mthd. * Kurt	.000	6	4E-005	21.14	< .001	.001
Clstrs * Clstr. Size	.024	4	.006	3094	< .001	.064
Clstrs * Corr	.016	4	.004	2138	< .001	.045
Clstrs * Skew	.006	6	.001	518.3	< .001	.017
Clstrs * Kurt	.003	6	.000	219.6	< .001	.007
Clstr. Size * Corr	.005	4	.001	664.9	< .001	.014
Clstr. Size * Skew	.009	6	.001	736.8	< .001	.024
Clstr. Size * Kurt	.007	6	.001	643.6	< .001	.021
Corr * Skew	.003	6	.001	272.0	< .001	.009
Corr * Kurt	.003	6	.001	300.8	< .001	.010
Skew * Kurt	.001	2	.001	319.0	< .001	.003
Error	.350	181786	2E-006			
Total	.534	181875				
Corrected Total	.488	181874				

a. R Squared = .284 (Adjusted R Squared = .284)

Note. The model eta squared for the full factorial 6-way model was .618 with df= 696. The model eta squared for only the main and the two-way interaction effects was .284 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .334(i.e., .618-.284).

Figure 4 displays the results of the post hoc confidence intervals investigating standard error bias among the cross-level interaction term. The findings were similar to results reported for the level-1 and level-2 standard errors. Namely, as the number of clusters increased, the amount of bias in the dependent variable decreased. Further, when kurtosis increased to four and seven and the correlation among variables increased to *r* = .70, the standard errors were negatively biased. This was especially true when the number of clusters was held at 30 and cluster size ranged from 10 to 30. However, bias in the cross-level interaction terms was less in magnitude when compared to bias in both the level-1 and level-2 standard errors. In these same conditions, the weighted least squares estimator produced less biased results.

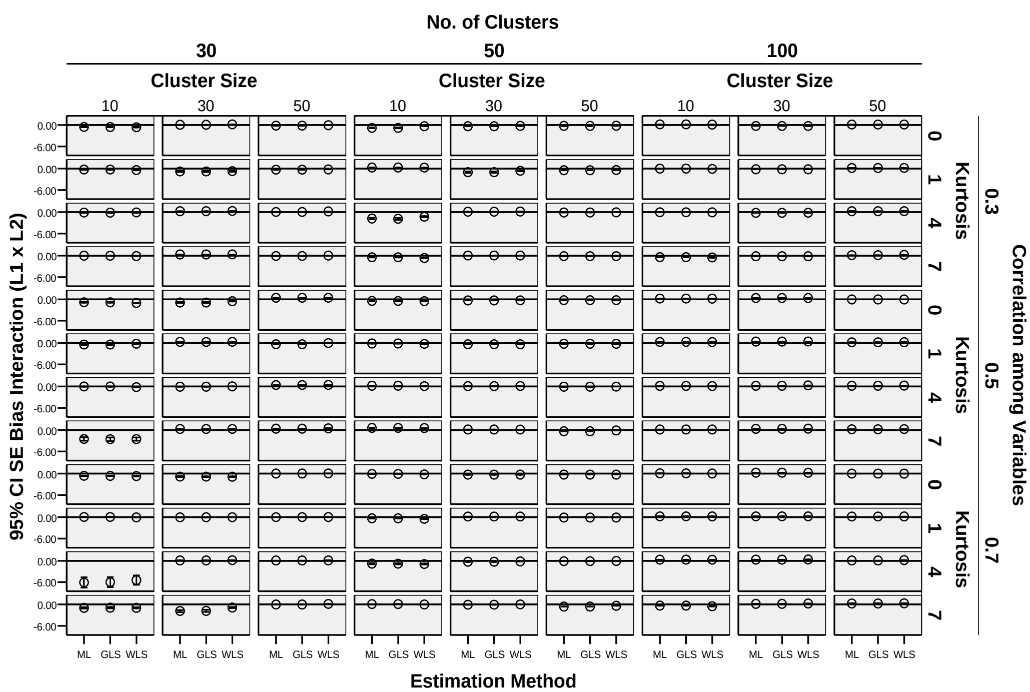

Figure 4. Confidence intervals examining bias among the cross-level interaction standard error, estimation method, number of clusters, cluster size, correlation among variables, and kurtosis.

Summary of Model One

The results from model one indicated that of the six conditions examined, cluster size and the correlation among variables had a statistically significant effect on bias in the level-1 parameter estimate and the cross-level interaction term, but explained little variation in the criterion variable. However, when investigating the standard errors, the antithesis was true. The number of clusters appeared to have the greatest impact on bias among the standard errors. When examining interaction effects via confidence intervals as a post hoc procedure, bias was most prevalent when the number of clusters and cluster size was smaller. Most notably, when the number of clusters was 30 and 50 and cluster size ranged from 10 to 50, the standard errors were biased downward. In this same setting, when kurtosis was severe (kurtosis = 7), the weighted least square estimator produced slightly less biased results when compared to the maximum likelihood and generalized least squares estimators, respectively.

However, as sample size increased, bias decreased with each estimator producing similar results when the number of clusters increased to 100 regardless of cluster size.

Due to the simplicity of the model discussed above, with only one predictor variable at both level-1 and level-2 and a cross-level interaction term between the level-1 and level-2 predictor variables, a more complex model was examined.

Model Two

The model investigated in this section included two level-1 and one level-2 predictor variables and a cross-level interaction term between the initial level-1 and the level-2 variable. The model is displayed below.

Level-1 equation $Y_{ij} = \beta_0 + \beta_1 X1_{ij} + \beta_2\ X4_{ij} + \varepsilon_{ij}$

Level-2 equation $\beta_0 = \gamma 00 + U0_j$

$\beta_1 = \gamma 10 + \gamma_{11} XC2BAR$

$\beta_2 = \gamma 20 + U_{1j}$

Combined equation $Y = \gamma_{00} + \gamma_0 X1 + \gamma_{11}\ X2BAR{*}X1 + \gamma_{20}\ x_4 + U_0 + \varepsilon$

Descriptive Measures of Bias among Parameter Estimates

Tables 16 and 17 display the descriptive results of the bias among level-1 and cross-level interaction parameter estimates by data normality, sample size, and estimation method. Mean bias was similar across estimation methods with the greatest amount of bias associated with the cross-level interaction term. Bias among estimation methods regarding the interaction term ranged from 1.22 (*SD* = 18.51) for the weighted least squares estimator in severely nonnormal data when the number of clusters equaled 50 and cluster size was 10 to 1.57 (*SD* = 7.63) for the maximum likelihood estimator in moderately nonnormal data when the number of clusters equaled 50 and cluster size was held at 30. Overall, as sample size increased, the amount of bias decreased. Further, when data nonnormality conditions increased from moderately to severely nonnormal, the weighted least squares estimator produced slightly less biased results in moderate to large sample sizes.

Table 16

Descriptive Results of Bias among Level-1 Parameter Estimates and Selected Design Conditions

		Estimation Method											
		ML				GLS				WLS			
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	.07	.24	.00	.11	.07	.25	.01	.11	.14	1.10	-.03	.55
	N = 900 (30 x 30)	.03	.04	-.01	.00	.03	.04	-.01	.00	.03	.04	-.01	.00
	N = 1500 (30 x 50)	.00	.00	.00	.01	.00	.00	.00	.01	.00	.00	.00	.01
	N = 500 (50 x 10)	.02	.05	.01	.03	.02	.04	.01	.03	.02	.04	.01	.03
	N = 1500 (50 x 30)	.01	.01	.00	.01	.01	.01	.00	.01	.01	.01	.00	.01
	N = 2500 (50 x 50)	-.01	.01	.00	.00	-.01	.01	.00	.00	-.01	.01	.00	.00
	N = 1000 (100 x 10)	.00	.01	-.01	.01	.00	.01	-.01	.01	.00	.01	-.01	.01
	N = 3000 (100 x 30)	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01
	N = 5000 (100 x 50)	.01	.02	.02	.04	.01	.02	.02	.04	.01	.02	.02	.04
Moderately Nonnormal	N = 300 (30 x 10)	.07	.31	.02	.17	.10	.43	.00	.22	.10	1.54	.00	.78
	N = 900 (30 x 30)	-.01	.04	.00	.01	-.01	.04	.00	.01	-.01	.04	.00	.01
	N = 1500 (30 x 50)	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01
	N = 500 (50 x 10)	.05	.13	.00	.03	.05	.13	.00	.03	.05	.13	.00	.04
	N = 1500 (50 x 30)	.00	.01	.01	.02	.00	.01	.01	.02	.00	.01	.01	.02
	N = 2500 (50 x 50)	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01
	N = 1000 (100 x 10)	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01
	N = 3000 (100 x 30)	.00	.00	.00	.01	.00	.00	.00	.01	.00	.00	.00	.01
	N = 5000 (100 x 50)	.01	.03	.03	.07	.01	.03	.03	.07	.01	.03	.03	.07
Severely Nonnormal	N = 300 (30 x 10)	.07	.16	.02	.09	.04	.26	.03	.14	.03	.39	.04	.20
	N = 900 (30 x 30)	.00	.05	.00	.01	.00	.05	.00	.01	.00	.05	.00	.01
	N = 1500 (30 x 50)	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01
	N = 500 (50 x 10)	.01	.08	.01	.02	.01	.08	.01	.02	.01	.23	.01	.02
	N = 1500 (50 x 30)	.00	.01	.02	.03	.00	.01	.02	.03	.00	.01	.02	.03
	N = 2500 (50 x 50)	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01
	N = 1000 (100 x 10)	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01	.00	.01
	N = 3000 (100 x 30)	.00	.00	.00	.01	.00	.00	.00	.01	.00	.00	.00	.01
	N = 5000 (100 x 50)	.01	.03	.02	.05	.01	.03	.02	.05	.01	.03	.02	.05

Table 17
Descriptive Results of Bias among Cross-Level Interaction Parameter Estimates and Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	.00	.35	-.03	.36	-.14	2.74
	N = 900 (30 x 30)	-.03	1.12	-.03	1.10	.08	1.10
	N = 1500 (30 x 50)	.00	.30	-.02	.31	.02	.32
	N = 500 (50 x 10)	-.08	5.00	.07	4.76	.15	4.95
	N = 1500 (50 x 30)	-.03	2.34	.00	2.31	-.13	2.32
	N = 2500 (50 x 50)	.01	1.60	.03	1.48	-.04	1.52
	N = 1000 (100 x 10)	-.05	1.47	.05	1.43	.06	1.42
	N = 3000 (100 x 30)	-.07	1.66	-.11	1.74	.02	1.75
	N = 5000 (100 x 50)	.00	.52	-.03	.49	.02	.49
Moderately Nonnormal	N = 300 (30 x 10)	.00	.40	.02	.39	-.10	3.96
	N = 900 (30 x 30)	-.17	2.01	-.15	1.96	-.16	2.02
	N = 1500 (30 x 50)	-.03	.30	-.03	.29	-.04	.30
	N = 500 (50 x 10)	.01	4.32	-.10	4.28	.16	6.65
	N = 1500 (50 x 30)	.18	2.73	.07	2.87	.09	2.80
	N = 2500 (50 x 50)	-.41	1.68	-.44	1.64	-.45	1.69
	N = 1000 (100 x 10)	-.17	1.46	-.20	1.45	-.16	1.54
	N = 3000 (100 x 30)	.06	1.66	.06	1.67	.06	1.67
	N = 5000 (100 x 50)	.27	1.69	.32	1.75	.30	1.72
Severely Nonnormal	N = 300 (30 x 10)	-.04	.47	-.02	.49	-.04	.62
	N = 900 (30 x 30)	-.40	2.56	-.40	2.56	-.39	2.58
	N = 1500 (30 x 50)	-.22	.63	-.22	.62	-.21	.63
	N = 500 (50 x 10)	-1.06	8.32	-1.08	8.13	-1.22	18.15
	N = 1500 (50 x 30)	1.57	7.63	1.54	7.66	1.49	7.68
	N = 2500 (50 x 50)	.13	2.14	.13	2.13	.18	2.17
	N = 1000 (100 x 10)	.88	3.97	.88	3.98	.78	3.95
	N = 3000 (100 x 30)	.63	2.66	.61	2.69	.55	2.71
	N = 5000 (100 x 50)	-.29	1.31	-.32	1.35	-.30	1.33

Similar to the results reported for the model-one standard errors, Tables 18 through 20 indicate that bias among the standard errors for each of the parameter estimates were similar in magnitude across estimation methods. However, the standard errors appeared to be negatively biased for each estimator in smaller sample sizes when the number of clusters ranged from 30 to 50. Average bias among the standard errors ranged from -5.15 (*SD* = 5.52) with the weighted least squares estimator when the number of clusters equaled 50 and cluster size equaled 10 to .29 (*SD* = .30) with the maximum likelihood estimator when the number of clusters equaled 50 and cluster size was 50 in severely nonnormal data. The largest amount of bias among the standard errors was associated with the level-1 standard errors, while the least amount of bias was associated with the cross-level interaction term. Comparing bias among parameter estimates to the standard errors, the direct opposite was found. More succinctly, bias among the standard errors was greater than bias associated with the parameter estimates. Further, as data nonnormality increased to moderately and severely nonnormal, the weighted least squares estimator produced slightly less biased standard errors in moderate sample sizes. As the sample size increased to 3,000 and 5,000 respectively, bias among standard errors decreased in magnitude with similar results reported across the three estimation methods. Note that the standard deviation associated with bias among the level-2 standard error was larger than the remaining standard errors, indicating that there was more variation in standard error bias between clusters than within clusters.

Table 18
Descriptive Results of Bias among Level-1 Standard Errors and Selected Design Conditions

		Estimation Method											
		ML				GLS				WLS			
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	-.93	2.13	-.93	2.13	-.84	2.10	-.83	2.09	-.72	1.86	-.66	1.76
	N = 900 (30 x 30)	-2.24	4.24	-2.23	4.24	-2.21	4.23	-2.21	4.22	-2.23	4.37	-1.89	3.52
	N = 1500 (30 x 50)	-.14	.41	-.14	.41	-.14	.41	-.13	.41	-.15	.42	-.14	.41
	N = 500 (50 x 10)	-4.89	5.06	-4.88	5.06	-4.94	5.13	-4.93	5.12	-5.15	5.52	-4.83	5.10
	N = 1500 (50 x 30)	-.29	.55	-.29	.55	-.29	.55	-.29	.55	-.27	.53	-.27	.53
	N = 2500 (50 x 50)	-.23	.56	-.23	.56	-.23	.56	-.23	.56	-.22	.55	-.23	.56
	N = 1000 (100 x 10)	.01	.37	.01	.37	.01	.37	.01	.37	.04	.31	.05	.30
	N = 3000 (100 x 30)	.00	.46	.00	.46	.00	.46	.00	.46	-.02	.47	-.01	.46
	N = 5000 (100 x 50)	-.20	.47	-.20	.47	-.20	.47	-.20	.47	-.20	.47	-.19	.46
Moderately Nonormal	N = 300 (30 x 10)	-1.00	1.60	-1.00	1.60	-.94	1.56	-.94	1.55	-.98	1.77	-.88	1.51
	N = 900 (30 x 30)	-.72	2.45	-.71	2.45	-.72	2.51	-.71	2.50	-.63	2.33	-.64	2.36
	N = 1500 (30 x 50)	-.64	1.39	-.64	1.39	-.64	1.39	-.64	1.39	-.57	1.30	-.59	1.34
	N = 500 (50 x 10)	-1.79	2.56	-1.79	2.55	-1.77	2.54	-1.77	2.54	-1.68	2.38	-1.67	2.41
	N = 1500 (50 x 30)	-.30	1.09	-.30	1.09	-.29	1.08	-.29	1.08	-.24	1.07	-.23	1.02
	N = 2500 (50 x 50)	-.26	.62	-.26	.62	-.26	.62	-.26	.62	-.23	.60	-.22	.57
	N = 1000 (100 x 10)	-.17	.76	-.16	.76	-.16	.76	-.16	.76	-.14	.71	-.13	.71
	N = 3000 (100 x 30)	-.07	.39	-.07	.39	-.07	.39	-.07	.39	-.05	.39	-.05	.39
	N = 5000 (100 x 50)	-.07	.57	-.07	.57	-.07	.57	-.07	.57	-.04	.53	-.04	.53
Severely Nonormal	N = 300 (30 x 10)	-1.29	2.23	-1.29	2.23	-1.20	2.31	-1.19	2.31	-1.15	2.08	-1.14	2.12
	N = 900 (30 x 30)	-1.50	3.82	-1.49	3.82	-1.47	3.78	-1.47	3.78	-1.22	3.62	-1.15	3.41
	N = 1500 (30 x 50)	-.38	.85	-.38	.85	-.38	.85	-.38	.85	-.20	.74	-.27	.80
	N = 500 (50 x 10)	-1.19	1.57	-1.19	1.57	-1.14	1.57	-1.14	1.57	-1.07	1.49	-.99	1.40
	N = 1500 (50 x 30)	-.13	.48	-.13	.48	-.13	.48	-.12	.48	-.03	.38	.00	.35
	N = 2500 (50 x 50)	-.07	.32	-.07	.32	-.07	.32	-.07	.32	.02	.26	.02	.24
	N = 1000 (100 x 10)	-.32	.82	-.31	.82	-.31	.81	-.31	.81	-.22	.69	-.19	.68
	N = 3000 (100 x 30)	-.08	.47	-.07	.47	-.08	.47	-.07	.47	.03	.39	-.01	.44
	N = 5000 (100 x 50)	.00	.49	.00	.49	.00	.50	.00	.50	.09	.41	.07	.45

Table 19
Descriptive Results of Bias among Level-2 Standard Error and Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	-1.04	1.79	-1.11	1.82	-1.03	1.67
	N = 900 (30 x 30)	-2.62	3.52	-2.62	3.50	-2.19	2.88
	N = 1500 (30 x 50)	.12	.25	.12	.25	.13	.23
	N = 500 (50 x 10)	-2.20	3.64	-2.23	3.64	-1.67	3.18
	N = 1500 (50 x 30)	-.50	.90	-.51	.90	-.35	.63
	N = 2500 (50 x 50)	.21	.44	.21	.44	.20	.45
	N = 1000 (100 x 10)	.23	.46	.24	.44	.23	.47
	N = 3000 (100 x 30)	.29	.14	.29	.14	.29	.14
	N = 5000 (100 x 50)	-.18	.43	-.18	.43	-.19	.44
Moderately Nonormal	N = 300 (30 x 10)	-.60	1.27	-.61	1.26	-.62	1.30
	N = 900 (30 x 30)	-.55	1.96	-.55	1.99	-.46	1.64
	N = 1500 (30 x 50)	-.29	.84	-.29	.85	-.23	.73
	N = 500 (50 x 10)	-.52	1.46	-.52	1.46	-.37	1.07
	N = 1500 (50 x 30)	-.67	1.79	-.67	1.79	-.64	1.81
	N = 2500 (50 x 50)	.19	.35	.19	.36	.18	.35
	N = 1000 (100 x 10)	.23	.48	.23	.48	.24	.43
	N = 3000 (100 x 30)	.22	.19	.22	.19	.22	.19
	N = 5000 (100 x 50)	.06	.26	.06	.26	.06	.26
Severely Nonormal	N = 300 (30 x 10)	-1.90	9.09	-2.24	10.04	-2.28	10.52
	N = 900 (30 x 30)	-1.66	4.24	-1.85	4.63	-1.80	4.41
	N = 1500 (30 x 50)	-.01	.48	-.01	.48	-.04	.51
	N = 500 (50 x 10)	-.72	2.39	-.89	2.81	-.96	2.99
	N = 1500 (50 x 30)	-.18	.65	-.19	.65	-.21	.64
	N = 2500 (50 x 50)	.29	.30	.29	.30	.29	.28
	N = 1000 (100 x 10)	-.31	2.04	-.31	2.04	-.28	1.76
	N = 3000 (100 x 30)	.24	.22	.24	.22	.22	.22
	N = 5000 (100 x 50)	.13	.18	.13	.18	.12	.19

Table 20
Descriptive Results of Bias among Interaction Standard Error and Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	-1.04	1.79	-1.11	1.82	-1.03	1.67
	N = 900 (30 x 30)	-1.62	1.52	-1.62	1.50	-1.19	1.88
	N = 1500 (30 x 50)	.12	.25	.12	.25	.13	.23
	N = 500 (50 x 10)	-1.20	1.64	-1.23	1.64	-1.67	1.18
	N = 1500 (50 x 30)	-.50	.90	-.51	.90	-.35	.63
	N = 2500 (50 x 50)	.21	.44	.21	.44	.20	.45
	N = 1000 (100 x 10)	.23	.46	.24	.44	.23	.47
	N = 3000 (100 x 30)	.29	.14	.29	.14	.29	.14
	N = 5000 (100 x 50)	-.18	.43	-.18	.43	-.19	.44
Moderately Nonormal	N = 300 (30 x 10)	-.60	1.27	-.61	1.26	-.62	1.30
	N = 900 (30 x 30)	-.55	1.96	-.55	1.99	-.46	1.64
	N = 1500 (30 x 50)	-.29	.84	-.29	.85	-.23	.73
	N = 500 (50 x 10)	-.52	1.46	-.52	1.46	-.37	1.07
	N = 1500 (50 x 30)	-.67	1.79	-.67	1.79	-.64	1.81
	N = 2500 (50 x 50)	.19	.35	.19	.36	.18	.35
	N = 1000 (100 x 10)	.23	.48	.23	.48	.24	.43
	N = 3000 (100 x 30)	.22	.19	.22	.19	.22	.19
	N = 5000 (100 x 50)	.06	.26	.06	.26	.06	.26
Severely Nonormal	N = 300 (30 x 10)	-1.90	1.09	-1.24	1.04	-1.28	1.52
	N = 900 (30 x 30)	-1.66	1.24	-1.85	1.63	-1.80	1.41
	N = 1500 (30 x 50)	-.01	.48	-.01	.48	-.04	.51
	N = 500 (50 x 10)	-.72	1.39	-.89	1.81	-.96	1.99
	N = 1500 (50 x 30)	-.18	.65	-.19	.65	-.21	.64
	N = 2500 (50 x 50)	.29	.30	.29	.30	.29	.28
	N = 1000 (100 x 10)	-.31	1.04	-.31	1.04	-.28	1.76
	N = 3000 (100 x 30)	.24	.22	.24	.22	.22	.22
	N = 5000 (100 x 50)	.13	.18	.13	.18	.12	.19

Bivariate Results

Table 21 reports -moment correlations between the bias among parameter estimates, their associated standard errors, and each of the six conditions examined in the study. Regarding bias among parameter estimates, each of the design conditions shared a slight statistically significant relation with bias among the initial level-1 parameter estimate. Note however, the effect size for each condition explained less than one percent of the variation in the outcome variable. As for bias among the subsequent level-1 parameter estimate, each of the design conditions was statistically significant, with the exception of the number of clusters. Similar to the initial level−1 parameter estimate, each of the correlations were indeed negligible, explaining less than one percent of the bias variance. None of the conditions were statistically significant when examining the effect of the design conditions on bias in the level-2 parameter estimate. Similar to the initial level−1 parameter estimate, the number of clusters and the correlation among variables shared a statistically significant association with bias among the cross-level interaction term, with each explaining less than one percent of the variance.

Concerning bias among standard errors, the largest correlation was noted between the level-2 parameter estimate standard error and the number of clusters (*r* = -.253, $p < .001$), indicating that as the number of clusters increased, the amount of bias associated with the level-2 standard error decreased. The 95% CI ranged from -.257 to -.248 with an effect size of .064, indicating that approximately six percent of the variance in bias among the level-2 parameter estimate's standard error could be explained by the number of clusters. Similar effects were noted for cluster size. In addition, skew shared a statistically significant positive relation with bias among the level–2 standard error (*r* = .046, $p < .01$). The 95% CI ranged from .041 to .051. However, the effect size was .002, indicating that skew among variables explained less than one percent of the variance in the outcome variable. Remaining statistically significant correlations among bias associated with the standard errors ranged from *r* = -.074, $p < .01$ between the cross-level interaction term and the number of clusters (95% CI ranged from -.079 to -.068) to *r* = .028, $p < .01$ between the initial level–1 parameter estimate's standard error and correlation among variables (95% CI ranged from .023 to .032).

Table 21

Pearson Product-Moment Correlations among Parameter Estimate and Standard Error Bias and Selected Design Conditions

	1	2	3	4	5	6	7	8	9	10	11	12	13
Bias Level-1 Predictor (1)	1.00												
Bias Level-1 Predictor (2)	-.034**	1.00											
Bias Level-2 Predictor (3)	-.003	.000	1.00										
Bias Cross-Level Interact.(4)	-.021**	.057**	-.013**	1.00									
Bias Level-1 Predictor SE (5)	.241**	-.239**	-.003	-.054**	1.00								
Bias Level-1 Predictor SE (6)	.240**	-.239**	-.003	-.052**	.030**	1.00							
Bias Level-2 Predictor SE (7)	-.024**	-.065**	.000	-.008**	.028**	.033**	1.00						
Bias Cross-Level Interact. SE (8)	.071**	-.098**	-.001	-.045**	.797**	.801**	.403**	1.00					
No. of Clusters (9)	-.030**	.004	-.002	.032**	-.010**	.003	-.253**	-.011**	1.00				
Cluster Size (10)	-.047**	.007**	-.003	-.001	-.007**	-.008**	-.051**	-.074**	-.004	1.00			
Correlation among Variables (11)	.027**	-.030**	.002	-.063**	.028**	.016**	.009**	.004	.000	.000	1.00		
Skew (12)	-.007**	.013**	-.001	.026**	.002	.003	.046**	.019**	-.002	-.002	.000	1.00	
Kurtosis (13)	-.022**	.024**	-.002	.002	-.009**	-.008**	-.003	-.004	-.001	-.002	.000	.403**	1.00

**. Correlation is significant at the 0.01 level (2-tailed). The sample size for each coefficient was 181,875.

Factorial ANOVA Results of Parameter Estimates

Tables 22 through 25 display the results of the six design conditions on the parameter estimates. Statistically significant main effects for the level-1 parameter estimates included cluster size, the number of clusters, correlation among variables skew, and kurtosis. Estimation method as a main effect was not statistically significant. However, the interaction terms that included estimation method by the number of clusters, estimation method by skew, and estimation method by kurtosis were statistically significant with each of the statistically significant effects explaining less than one percent of the variance in the outcome variable. Similar results were found for both the initial and subsequent level-1 parameter estimates.

In contrast, the impact of the study’s design conditions on bias among the level-2 parameter estimate was negligible. More succinctly, none of the conditions investigated shared a statistically significant relation with bias among the level-2 parameter estimate.

The results displayed for the cross-level interaction parameter estimate revealed that the estimation method was not statistically significant either as a main or interaction effect. Although the remaining design conditions were statistically significant, the effect size, or amount of variation in bias explained by the remaining design conditions, was less than or equal to one percent (partial η^2 ranged from .001 to .010).

Table 22

Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Initial Level-1 Parameter Estimate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	22713911.70	88	258113	77.24	< .001	.036[a]
Intercept	1106379.024	1	1E+006	331.1	< .001	.002
Est. Mthd	2636.637	2	1318.3	.395	.674	.000
Clstrs	187454.943	2	93727	28.05	< .001	.000
Clstr. Size	1449622.981	2	724811	216.9	< .001	.002
Corr	852893.755	2	426447	127.6	< .001	.001
Skew	432239.979	3	144080	43.12	< .001	.001
Kurt	619427.272	3	206476	61.79	< .001	.001
Est. Mthd. * Clstrs	1859.269	4	464.82	.139	.968	.000
Est. Mthd. * Clstr. Size	1294.692	4	323.67	.097	.983	.000
Est. Mthd. * Corr	33885.740	4	8471.4	2.535	.038	.000
Est. Mthd. * Skew	54779.332	6	9129.9	2.732	.012	.000
Est. Mthd. * Kurt	61538.690	6	10256	3.069	.005	.000
Clstrs * Clstr. Size	2653949.588	4	663487	198.5	< .001	.004
Clstrs * Corr	749552.033	4	187388	56.08	< .001	.001
Clstrs * Skew	2592145.419	6	432024	129.3	< .001	.004
Clstrs * Kurt	3750035.338	6	625006	187.0	< .001	.006
Clstr. Size * Corr	3122976.584	4	780744	233.6	< .001	.005
Clstr. Size * Skew	1667345.534	6	277891	83.16	< .001	.003
Clstr. Size * Kurt	1437729.164	6	239622	71.71	< .001	.002
Corr * Skew	2279217.466	6	379870	113.7	< .001	.004
Corr * Kurt	1005010.383	6	167502	50.12	< .001	.002
Skew * Kurt	46579.760	2	23290	6.969	.001	.000
Error	606627979.0	181533	3341.7			
Total	630860078.3	181622				
Corrected Total	629341890.7	181621				

a. R Squared = .036 (Adjusted R Squared = .036)

Note. The model eta squared for the full factorial 6-way model was .121 with df=696. The model eta squared for only the main and the two-way interaction effects was .036 with df=88. Thus, the eta squared effect size for all the unreported two-three-, four-, five-, and the six-way interaction effects was .085(i.e., .121-.036).

Table 23
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Subsequent Level-1 Parameter Estimate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	6089571.4	88	69200	81.53	< .001	.038[a]
Intercept	493594.581	1	493595	581.5	< .001	.003
Est. Mthd	665.880	2	332.94	.392	.676	.000
Clstrs	19434.863	2	9717.4	11.45	< .001	.000
Clstr. Size	20096.990	2	10048	11.84	< .001	.000
Corr	98082.563	2	49041	57.78	< .001	.001
Skew	228291.585	3	76097	89.66	< .001	.001
Kurt	300048.325	3	100016	117.8	< .001	.002
Est. Mthd. * Clstrs	305.090	4	76.273	.090	.986	.000
Est. Mthd. * Clstr. Size	321.454	4	80.364	.095	.984	.000
Est. Mthd. * Corr	8573.147	4	2143.3	2.525	.039	.000
Est. Mthd. * Skew	14352.520	6	2392.1	2.818	.010	.000
Est. Mthd. * Kurt	15477.891	6	2579.6	3.039	.006	.000
Clstrs * Clstr. Size	762337.248	4	190584	224.5	< .001	.005
Clstrs * Corr	642856.546	4	160714	189.3	< .001	.004
Clstrs * Skew	603661.637	6	100610	118.5	< .001	.004
Clstrs * Kurt	599630.269	6	99938	117.7	< .001	.004
Clstr. Size * Corr	679236.595	4	169809	200.1	< .001	.004
Clstr. Size * Skew	757233.298	6	126206	148.7	< .001	.005
Clstr. Size * Kurt	752485.594	6	125414	147.8	< .001	.005
Corr * Skew	330111.540	6	55019	64.82	< .001	.002
Corr * Kurt	457453.492	6	76242	89.83	< .001	.003
Skew * Kurt	41873.353	2	20937	24.67	< .001	.000
Error	154079928	181533	848.77			
Total	160584402	181622				
Corrected Total	160169499	181621				

a. R Squared = .038 (Adjusted R Squared = .038)

Note. The model eta squared for the full factorial 6-way model was .124 with df=.696. The model eta squared for only the main and the two-way interaction effects was .038 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .086(i.e., .124-.038).

Table 24
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Level-2 Parameter Estimate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2678146	88	30433	.861	.821	.000[a]
Intercept	.005	1	.005	.000	1.000	.000
Est. Mthd	6889.414	2	3444.7	.097	.907	.000
Clstrs	1658.468	2	829.23	.023	.977	.000
Clstr. Size	155.446	2	77.723	.002	.998	.000
Corr	25863.00	2	12932	.366	.694	.000
Skew	91391.80	3	30464	.861	.460	.000
Kurt	179792.3	3	59931	1.695	.166	.000
Est. Mthd. * Clstrs	146839.4	4	36710	1.038	.386	.000
Est. Mthd. * Clstr. Size	162166.0	4	40541	1.146	.333	.000
Est. Mthd. * Corr	62352.75	4	15588	.441	.779	.000
Est. Mthd. * Skew	172213.7	6	28702	.812	.561	.000
Est. Mthd. * Kurt	337513.8	6	56252	1.591	.145	.000
Clstrs * Clstr. Size	209641.6	4	52410	1.482	.205	.000
Clstrs * Corr	44124.70	4	11031	.312	.870	.000
Clstrs * Skew	280349.6	6	46725	1.321	.244	.000
Clstrs * Kurt	335634.9	6	55939	1.582	.148	.000
Clstr. Size * Corr	56918.29	4	14230	.402	.807	.000
Clstr. Size * Skew	161299.3	6	26883	.760	.601	.000
Clstr. Size * Kurt	320512.2	6	53419	1.510	.170	.000
Corr * Skew	272839.8	6	45473	1.286	.260	.000
Corr * Kurt	87390.92	6	14565	.412	.872	.000
Skew * Kurt	1296.581	2	648.29	.018	.982	.000
Error	6.4E+009	181533	35366			
Total	6.4E+009	181622				
Corrected Total	6.4E+009	181621				

a. R Squared = .000 (Adjusted R Squared = .000)

Note. The model eta squared for the full factorial 6-way model was .003 with df=696. The model eta squared for only the main and the two-way interaction effects was .000 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .003(i.e., .003-.000).

Table 25

Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Cross-Level *Interaction Parameter Estimate*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	186572.977	88	2120.147	142.870	< .001	.065[a]
Intercept	215.789	1	215.789	14.541	< .001	.000
Est. Mthd	14.862	2	7.431	.501	.606	.000
Clstrs	3504.385	2	1752.192	118.075	< .001	.001
Clstr. Size	4118.907	2	2059.453	138.780	< .001	.002
Corr	18714.261	2	9357.131	630.547	< .001	.007
Skew	1746.913	3	582.304	39.240	< .001	.001
Kurt	5771.225	3	1923.742	129.635	< .001	.002
Est. Mthd. * Clstrs	11.115	4	2.779	.187	.945	.000
Est. Mthd. * Clstr. Size	9.123	4	2.281	.154	.961	.000
Est. Mthd. * Corr	8.705	4	2.176	.147	.965	.000
Est. Mthd. * Skew	46.776	6	7.796	.525	.790	.000
Est. Mthd. * Kurt	50.205	6	8.367	.564	.759	.000
Clstrs * Clstr. Size	12820.025	4	3205.006	215.975	< .001	.005
Clstrs * Corr	27105.038	4	6776.260	456.630	< .001	.010
Clstrs * Skew	6071.926	6	1011.988	68.195	< .001	.002
Clstrs * Kurt	3221.495	6	536.916	36.181	< .001	.001
Clstr. Size * Corr	19991.685	4	4997.921	336.794	< .001	.007
Clstr. Size * Skew	8797.246	6	1466.208	98.803	< .001	.003
Clstr. Size * Kurt	10142.320	6	1690.387	113.910	< .001	.004
Corr * Skew	25542.224	6	4257.037	286.868	< .001	.009
Corr * Kurt	18258.569	6	3043.095	205.064	< .001	.007
Skew * Kurt	6950.790	2	3475.395	234.196	< .001	.003
Error	2693897.388	181533	14.840			
Total	2880687.245	181622				
Corrected Total	2880470.365	181621				

a. R Squared = .065 (Adjusted R Squared = .064)

Note. The model eta squared for the full factorial 6-way model was .262 with df= 696. The model eta squared for only the main and the two-way interaction effects was .065 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .197(i.e., .262-.065).

Factorial ANOVA Results of Standard Errors

Table 26 reports results of the impact of the six design conditions on bias among the initial level-1 parameter estimate’s standard error. Each of the main effects was statistically significant with the number of clusters and cluster size explaining the greatest amount of variation in the outcome variable (partial η^2=.024). In addition, the estimation method was statistically significant, but explained little variation in the dependent variable. Regarding the interaction terms, the number of clusters by cluster size explained the greatest amount of variance in bias among the initial level-1 standard error (partial η^2=.046) while the number of clusters by kurtosis explained approximately three percent of the variance in the criterion variable (partial η^2 =.026). The post hoc analysis results of the statistically significant effects are displayed in Figure 5.

Table 26

Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Initial Level-1 Standard Error

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	126653.432	88	1439.244	558.837	< .001	.213[a]
Intercept	39646.357	1	39646.357	15394.085	< .001	.078
Est. Mthd	215.569	2	107.784	41.851	< .001	.000
Clstrs	10799.017	2	5399.509	2096.548	< .001	.023
Clstr. Size	8900.045	2	4450.023	1727.877	< .001	.019
Corr	3620.960	2	1810.480	702.982	< .001	.008
Skew	4186.806	3	1395.602	541.891	< .001	.009
Kurt	2002.905	3	667.635	259.233	< .001	.004
Est. Mthd. * Clstrs	23.367	4	5.842	2.268	.059	.000
Est. Mthd. * Clstr. Size	14.484	4	3.621	1.406	.229	.000
Est. Mthd. * Corr	27.312	4	6.828	2.651	.031	.000
Est. Mthd. * Skew	24.245	6	4.041	1.569	.152	.000
Est. Mthd. * Kurt	30.990	6	5.165	2.006	.061	.000
Clstrs * Clstr. Size	22297.525	4	5574.381	2164.449	< .001	.046
Clstrs * Corr	10183.331	4	2545.833	988.509	< .001	.021
Clstrs * Skew	2311.220	6	385.203	149.569	< .001	.005
Clstrs * Kurt	12632.395	6	2105.399	817.495	< .001	.026
Clstr. Size * Corr	6639.851	4	1659.963	644.539	< .001	.014
Clstr. Size * Skew	4026.438	6	671.073	260.568	< .001	.009
Clstr. Size * Kurt	5367.621	6	894.603	347.361	< .001	.011
Corr * Skew	3884.246	6	647.374	251.366	< .001	.008
Corr * Kurt	1269.796	6	211.633	82.174	< .001	.003
Skew * Kurt	2149.131	2	1074.565	417.238	< .001	.005
Error	467525.164	181533	2.575			
Total	652885.709	181622				
Corrected Total	594178.597	181621				

a. R Squared = .213 (Adjusted R Squared = .213)

Note. The model eta squared for the full factorial 6-way model was .535 with df=696. The model eta squared for only the main and the two-way interaction effects was .213 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .322(i.e., .535-.213).

Figure 5 shows that in smaller sample sizes, the standard error associated with the initial level-1 parameter estimate was negatively biased regardless of data normality conditions (i.e., number of clusters = 30 and 50 and cluster size = 10). As data nonnormality increased from moderate to severe, the weighted least squares estimator produced slightly less biased standard errors in moderately large samples (when the number of clusters was 30 and 50 and cluster size ranged from 30 to 50). When the number of clusters was held at 100, the estimation methods produced similar results with the weighted least squares estimator producing slightly less biased standard errors for the initial level-1 parameter estimate, especially when kurtosis was severe (kurtosis = 7).

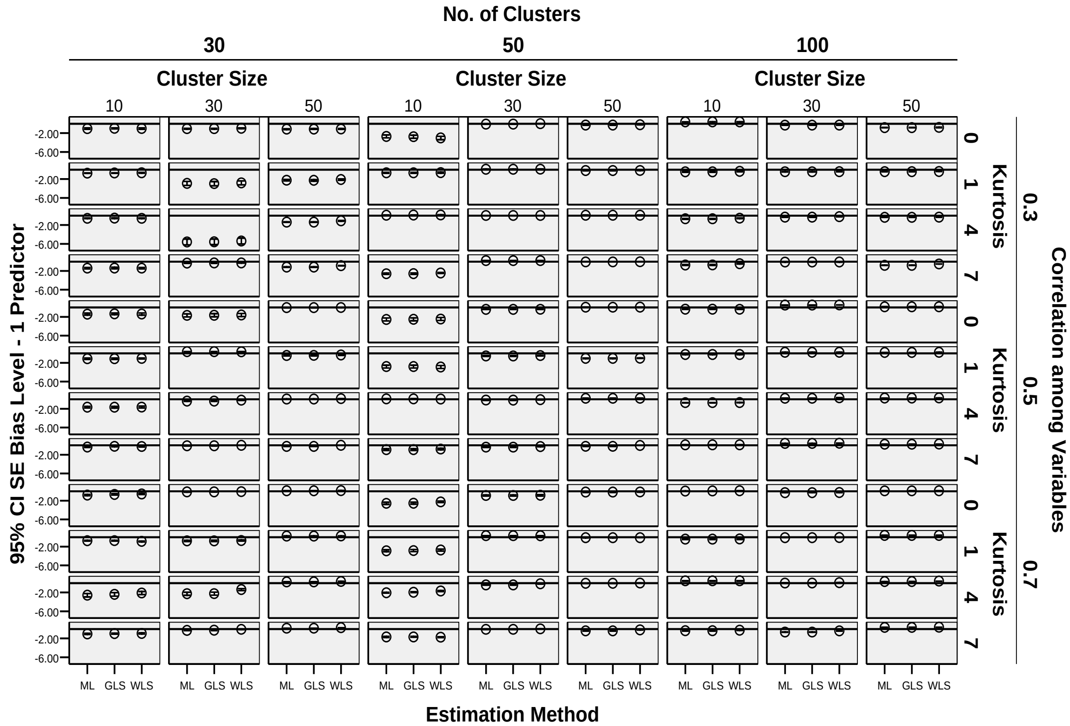

Figure 5. Confidence intervals examining bias among the initial level-1 standard error, estimation method, number of clusters, cluster size, correlation among variables, and kurtosis.

Analogous to the initial level-1 parameter estimate's standard error, the results displayed in Table 27 indicate that among the main effects, the number of clusters explained the greatest amount of variation in the subsequent level-1 parameter estimate's standard error. Although the estimation method was statistically significant, the estimation method explained less than one percent of the variance in the outcome variable. As for interaction effects, cluster by cluster size explained approximately five percent (partial $\eta^2 = .046$) of the variance while cluster size by kurtosis explained approximately three percent of the bias variance (η^2=.027) among the subsequent level-1 parameter estimate's standard error.

The resulting post hoc analysis displayed in Figure 6 revealed that in small sample sizes, the standard errors were negatively biased. As the number of clusters increased to 100, however, bias was negligible with the weighted least squares estimator producing slightly less biased standard errors in moderately and severely nonnormal conditions (kurtosis = 1-7) in moderately large sample sizes.

Table 27

Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Subsequent Level-1 Standard Error

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	123664.681	88	1405.3	565.4	< .001	.215[a]
Intercept	38830.362	1	38830	15622	< .001	.079
Est. Mthd	254.118	2	127.06	51.12	< .001	.001
Clstrs	10536.666	2	5268.3	2119	< .001	.023
Clstr. Size	8367.412	2	4183.7	1683	< .001	.018
Corr	3673.760	2	1836.9	739.0	< .001	.008
Skew	3981.575	3	1327.2	533.9	< .001	.009
Kurt	1786.207	3	595.40	239.5	< .001	.004
Est. Mthd. * Clstrs	46.972	4	11.743	4.724	.001	.000
Est. Mthd. * Clstr. Size	29.791	4	7.448	2.996	.017	.000
Est. Mthd. * Corr	16.643	4	4.161	1.674	.153	.000
Est. Mthd. * Skew	23.388	6	3.898	1.568	.152	.000
Est. Mthd. * Kurt	28.726	6	4.788	1.926	.073	.000
Clstrs * Clstr. Size	21684.220	4	5421.1	2181	< .001	.046
Clstrs * Corr	10403.347	4	2600.8	1046	< .001	.023
Clstrs * Skew	2121.428	6	353.57	142.2	< .001	.005
Clstrs * Kurt	12301.325	6	2050.2	824.8	< .001	.027
Clstr. Size * Corr	6990.164	4	1747.5	703.1	< .001	.015
Clstr. Size * Skew	4003.073	6	667.18	268.4	< .001	.009
Clstr. Size * Kurt	5227.697	6	871.28	350.5	< .001	.011
Corr * Skew	3670.895	6	611.82	246.1	< .001	.008
Corr * Kurt	1226.175	6	204.36	82.22	< .001	.003
Skew * Kurt	2057.974	2	1029.0	414.0	< .001	.005
Error	451228.17C	181533	2.486			
Total	632094.518	181622				
Corrected Total	574892.851	181621				

a. R Squared = .215 (Adjusted R Squared = .215)

Note. The model eta squared for the full factorial 6-way model was .535 with df=696. The model eta squared for only the main and the two-way interaction effects was .215 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .320(i.e., .535-.215).

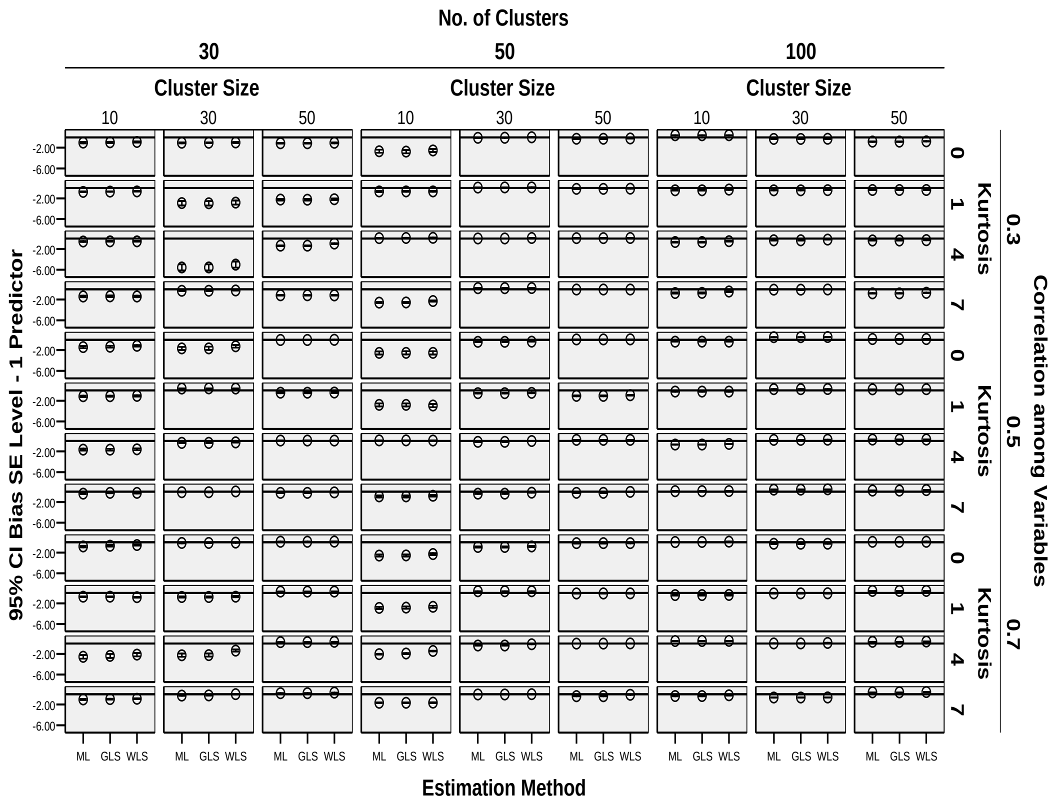

Figure 6. Confidence intervals examining bias among the subsequent level-1 standard error, estimation method, number of clusters, cluster size, correlation among variables, and kurtosis.

Regarding the impact of the main effect of the six design conditions on bias among the level-2 parameter estimate's standard error, the results reported in Table 28 reveal that the estimation method was statistically significant, which was similar to the findings from the level-1 standard errors. In addition, the remaining main effects were statistically significant with the number of clusters explaining the greatest amount of bias variance in the level-2 standard error (η^2=.058), while cluster size explained approximately one percent of the variance (η^2=.006). Further, the interaction terms, which included the estimation method by cluster size and estimation method by kurtosis, were statistically significant, but explained little variation in the outcome variable. Remaining interaction terms that included the number of clusters, cluster size, and data normality conditions were statistically significant and explained similar amounts of variance in the outcome variable (albeit small in magnitude).

The resulting post-hoc confidence intervals displayed in Figure 7 were similar to those reported for standard error bias among the level-1 standard errors. In smaller sample sizes, particularly when the number of clusters ranged from 10 to 50 and cluster size ranged from 30 to 50, the level-2 standard error was negatively biased. Moreover, in these same clusters the amount of negative bias increased as data nonnormality increased. As the number of clusters increased to 100, the estimation methods converged, producing similar results. Note that although the results were similar across estimation methods, when kurtosis and correlation among variables increased, the standard errors trended in a downward direction in small sample sizes (i.e., number of clusters = 30 to 50 and cluster size ranged from 10 to 30).

Table 28

Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Level-2 Standard Error

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	16003.07	88	181.85	833.5	< .001	.288[a]
Intercept	4752.699	1	4752.7	21784	< .001	.107
Est. Mthd	7.534	2	3.767	17.27	< .001	.000
Clstrs	2417.506	2	1208.8	5540	< .001	.058
Clstr. Size	238.926	2	119.46	547.6	< .001	.006
Corr	135.662	2	67.831	310.9	< .001	.003
Skew	117.068	3	39.023	178.9	< .001	.003
Kurt	136.963	3	45.654	209.3	< .001	.003
Est. Mthd. * Clstrs	.418	4	.104	.479	.751	.000
Est. Mthd. * Clstr. Size	2.407	4	.602	2.759	.026	.000
Est. Mthd. * Corr	.284	4	.071	.326	.861	.000
Est. Mthd. * Skew	2.464	6	.411	1.882	.080	.000
Est. Mthd. * Kurt	4.260	6	.710	3.254	.003	.000
Clstrs * Clstr. Size	4119.661	4	1029.9	4721	< .001	.094
Clstrs * Corr	2693.886	4	673.47	3087	< .001	.064
Clstrs * Skew	78.032	6	13.005	59.61	< .001	.002
Clstrs * Kurt	385.060	6	64.177	294.2	< .001	.010
Clstr. Size * Corr	1333.801	4	333.45	1528	< .001	.033
Clstr. Size * Skew	279.596	6	46.599	213.6	< .001	.007
Clstr. Size * Kurt	948.576	6	158.10	724.6	< .001	.023
Corr * Skew	238.967	6	39.828	182.5	< .001	.006
Corr * Kurt	522.294	6	87.049	399.0	< .001	.013
Skew * Kurt	108.962	2	54.481	249.7	< .001	.003
Error	39606.29	181533	.218			
Total	61138.28	181622				
Corrected Total	55609.36	181621				

a. R Squared = .288 (Adjusted R Squared = .287)

Note. The model eta squared for the full factorial 6-way model was .580 with df=696. The model eta squared for only the main and the two-way interaction effects was .288 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .292(i.e., .580-.288).

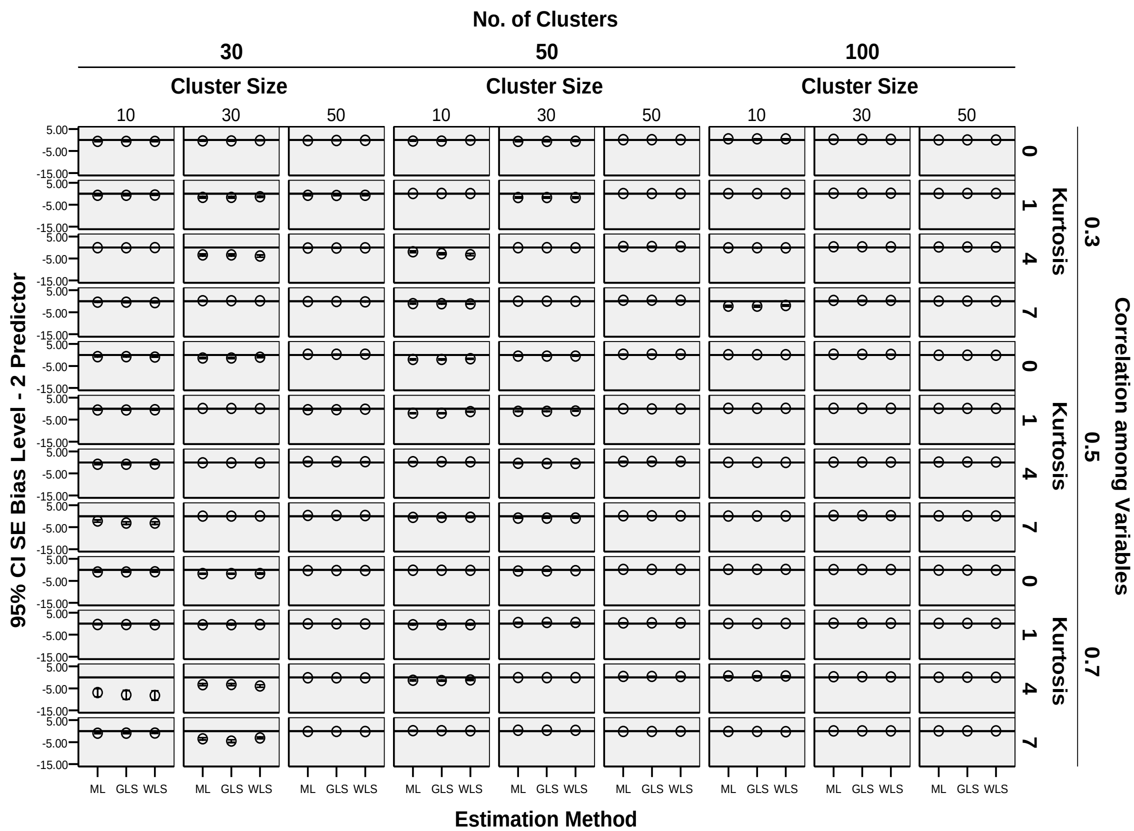

Figure 7. Confidence intervals examining bias among the level-2 standard error, estimation method, number of clusters, cluster size, correlation among variables, and kurtosis.

Table 29 reports the effect of the six design conditions on the cross-level interaction term's standard error bias. Each of the main effects was statistically significant, with each explaining approximately one percent of the variance in the outcome variable. As for interaction terms, each of the interaction terms were statistically significant with the exception of estimation method by the correlation among variables. In addition, the number of clusters by cluster size explained the largest amount of variation in the criterion variables (partial $\eta^2 = .053$), while the number of clusters by correlation among variables explained approximately three percent of the variance in bias associated with the cross-level interaction standard error. The remaining statistically significant interaction effects explained less than one percent of the variation in the outcome variable.

Table 29
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Cross-level Interaction Standard Error

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	41339.009	88	469.761	269.479	< .001	.116[a]
Intercept	3434.317	1	3434.317	1970.097	< .001	.011
Est. Mthd	42.974	2	21.487	12.326	< .001	.000
Clstrs	622.749	2	311.375	178.620	< .001	.002
Clstr. Size	1240.336	2	620.168	355.760	< .001	.004
Corr	3127.302	2	1563.651	896.989	< .001	.010
Skew	172.973	3	57.658	33.075	< .001	.001
Kurt	209.426	3	69.809	40.046	< .001	.001
Est. Mthd. * Clstrs	22.543	4	5.636	3.233	.012	.000
Est. Mthd. * Clstr. Size	60.807	4	15.202	8.720	< .001	.000
Est. Mthd. * Corr	7.562	4	1.891	1.084	.362	.000
Est. Mthd. * Skew	27.176	6	4.529	2.598	.016	.000
Est. Mthd. * Kurt	23.195	6	3.866	2.218	.038	.000
Clstrs * Clstr. Size	17739.617	4	4434.904	2544.084	< .001	.053
Clstrs * Corr	10917.401	4	2729.350	1565.693	< .001	.033
Clstrs * Skew	58.553	6	9.759	5.598	< .001	.000
Clstrs * Kurt	588.706	6	98.118	56.285	< .001	.002
Clstr. Size * Corr	1492.140	4	373.035	213.992	< .001	.005
Clstr. Size * Skew	372.061	6	62.010	35.572	< .001	.001
Clstr. Size * Kurt	1370.389	6	228.398	131.021	< .001	.004
Corr * Skew	289.143	6	48.191	27.645	< .001	.001
Corr * Kurt	659.695	6	109.949	63.072	< .001	.002
Skew * Kurt	82.485	2	41.242	23.659	< .001	.000
Error	316452.363	181533	1.743			
Total	361984.504	181622				
Corrected Total	357791.371	181621				

a. R Squared = .116 (Adjusted R Squared = .115)

Note. The model eta squared for the full factorial 6-way model was .204 with df=696. The model eta squared for only the main and the two-way interaction effects was .116 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .088(i.e., .204-.116).

Figure 8 displays the post hoc results examining bias associated with the cross-level interaction term's standard error by the number of clusters, cluster size, correlation among variables, and estimation method. In smaller sample sizes, particularly when the number of clusters was 30 and 50 and cluster size ranged from 10 to 30, the standard errors were negatively biased. Further, when cluster size increased to 50 in the same number of clusters, the amount of bias decreased with the weighted least squares estimator producing less biased results as the correlation among variables increased. When cluster size was 100, each estimation method produced slight positive bias when cluster size was 10 and 30 and the correlation among variables ranged from *r* = .50 to *r* = .70. In comparison, when the number of clusters was held at 100 and cluster size at 50, each estimator produced less biased standard errors as the correlation among variables increased.

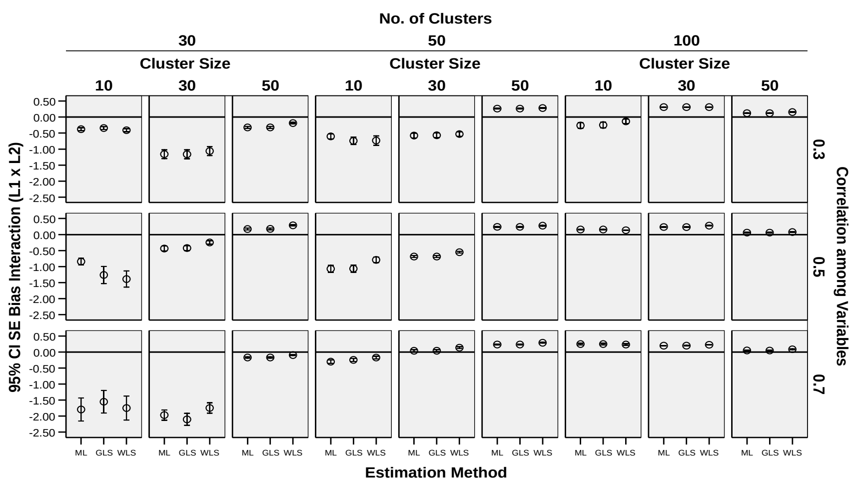

Figure 8. Confidence intervals examining bias among the cross-level interaction standard error, estimation method, number of clusters, cluster size, and correlation among variables.

Summary of Model Two

The results from model two indicated that of the six conditions examined, the correlation, skew and kurtosis among variables had a statistically significant effect on bias among the initial level-1 parameter estimate while the correlation among variables had a statistically significant impact on the subsequent level-1 parameter estimate. However, when investigating the post-hoc findings, the results suggested that statistically significant effects reported for the level-1 parameter estimates were due primarily to the large sample size, as the confidence intervals associated with each condition examined did indeed include zero. Similar results were found for the cross-level interaction parameter estimates.

The number of clusters appeared to have the greatest impact on bias among the standard errors. When examining interaction effects via confidence intervals as a post hoc procedure, bias was most prevalent across the standard errors investigated when the number of clusters and cluster size was smaller. This was most noted when the number of clusters was held at 30 and 50 and cluster size ranged from 10 to 30, where the standard errors were biased downward. In this same setting, when kurtosis ranged from one to seven, the weighted least square estimator produced slightly less biased results when compared to the maximum likelihood and generalized least squares estimators, respectively. However, as sample size increased, bias decreased with each estimator producing similar results when the number of clusters increased to 100 regardless of cluster size.

Due to the simplicity of the models discussed earlier, a more complex model was examined to determine whether three differing estimation methods in multilevel SEM, under varying conditions of data normality, and across different sample sizes produced more robust parameter estimates and standard errors.

Model Three

The model investigated in this section included two level-1, two level-2 predictor variables, and two cross-level interaction terms. The initial cross-level interaction term was comprised of the initial level-1 and the initial level-2 variables, while the subsequent interaction term included the subsequent level-1 and level-2 predictor variables. The model is displayed below.

Level-1 equation $Y_{ij} = \beta_0 + \beta_1 X1_{ij} + \beta_2\ X4_{ij} + \varepsilon_{ij}$

Level-2 equation $\beta_0 = \gamma 00 + U0_j$

$\beta_1 = \gamma 10 + \gamma_{11} XC2BAR$

$\beta_2 = \gamma 20 + \gamma 21\ XC5BAR$

Combined equation $Y = \gamma_{00} + \gamma_0 X1 + \gamma_{11}\ X2BAR{*}X1 + \gamma_{20}\ x_4 + \gamma_{21}{*}X5BAR{*}X_4\ U_0 + \varepsilon$

Descriptive Measures of Bias among Parameter Estimates

Tables 30 and 31 display the descriptive measures of the level-1 and cross-level interaction parameter estimates by data normality, sample size, and estimation method. The results revealed that mean bias among the level-1 parameter estimates was similar across estimation methods in each of the conditions examined. The standard deviations for both the maximum likelihood and generalized least squares methods were similar, while the weighted least squares estimator produced a larger standard deviation in small sample size, especially when the number of clusters equaled 30 and cluster size was 10.

Regarding bias among the cross-level interaction terms, the results were similar among the maximum likelihood and generalized least squares estimators, while the weighted least squares estimator produced results that were inconsistent in small sample sizes under normal data conditions based on the standard deviation. When the data increased in departure from normality, the weighted least squares estimator produced less biased results in small and moderate sample sizes. As sample size increased, especially when the number of clusters equaled 100 and cluster size was at least 50, the estimation methods converged, with each estimation method producing similar results.

Table 30

Descriptive Results of Bias among Level-1 Parameter Estimates and Selected Study Design Conditions

		Estimation Method											
		ML				GLS				WLS			
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Normal	N = 300 (30 X 10)	.005	.015	-.003	.007	.005	.016	-.003	.008	.009	.150	-.006	.075
	N = 900 (30 X 30)	-.002	.004	.007	.016	-.002	.004	.007	.016	-.002	.005	.007	.017
	N = 1500 (30 X 50)	.023	.027	-.106	.189	.023	.027	-.106	.188	.023	.027	-.107	.190
	N = 500 (50 X 10)	.003	.004	-.011	.022	.003	.004	-.012	.023	.003	.006	-.011	.029
	N = 1500 (50 X 30)	.006	.006	-.008	.007	.006	.006	-.008	.007	.006	.006	-.008	.007
	N = 2500 (50 X 50)	-.001	.004	.004	.006	-.001	.004	.004	.006	-.001	.004	.004	.006
	N = 1000 (100 X 10)	-.013	.012	.006	.007	-.013	.012	.006	.007	-.013	.012	.006	.007
	N = 3000 (100 X 30)	-.003	.009	.008	.011	-.003	.009	.008	.011	-.003	.009	.008	.011
	N = 5000 (100 X 50)	-.004	.005	.008	.007	-.004	.005	.008	.007	-.004	.005	.008	.007
Moderately Nonnormal	N = 300 (30 X 10)	-.006	.009	.002	.008	-.006	.008	.002	.008	-.006	.057	.001	.054
	N = 900 (30 X 30)	-.001	.006	.005	.012	-.001	.006	.005	.013	-.001	.010	.005	.020
	N = 1500 (30 X 50)	.008	.031	-.066	.253	.008	.031	-.066	.253	.008	.031	-.067	.261
	N = 500 (50 X 10)	.000	.009	.018	.049	.001	.012	.016	.049	.001	.033	.017	.058
	N = 1500 (50 X 30)	-.004	.009	-.004	.012	-.004	.009	-.004	.012	-.004	.009	-.004	.012
	N = 2500 (50 X 50)	-.004	.007	.000	.006	-.004	.007	.000	.006	-.004	.007	.000	.006
	N = 1000 (100 X 10)	-.002	.010	-.003	.007	-.002	.010	-.003	.007	-.002	.010	-.003	.007
	N = 3000 (100 X 30)	.004	.013	-.001	.015	.004	.013	-.001	.015	.004	.013	-.001	.015
	N = 5000 (100 X 50)	.000	.004	.004	.008	.000	.004	.004	.008	.000	.004	.004	.008
Severely Nonnormal	N = 300 (30 X 10)	.001	.009	-.002	.007	.001	.010	-.002	.007	.001	.163	-.002	.127
	N = 900 (30 X 30)	.000	.004	.003	.010	.000	.004	.003	.010	.000	.004	.003	.011
	N = 1500 (30 X 50)	.005	.018	-.251	.486	.005	.018	-.252	.487	.005	.018	-.251	.487
	N = 500 (50 X 10)	.001	.007	-.009	.058	.001	.007	-.009	.058	.000	.015	-.003	.276
	N = 1500 (50 X 30)	.001	.010	-.006	.009	.001	.010	-.007	.009	.001	.010	-.007	.009
	N = 2500 (50 X 50)	-.005	.007	.001	.006	-.005	.007	.001	.006	-.005	.007	.001	.006
	N = 1000 (100 X 10)	.001	.007	-.005	.007	.001	.007	-.005	.007	.001	.007	-.005	.007
	N = 3000 (100 X 30)	.011	.019	.004	.010	.011	.019	.004	.010	.011	.019	.004	.010
	N = 5000 (100 X 50)	.000	.004	.003	.008	.000	.004	.003	.008	.000	.004	.003	.008

Table 31

Descriptive Results of Bias among Cross-Level Interaction Parameter Estimates and Selected Design Conditions

		Estimation Method											
		ML				GLS				WLS			
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Normal	N = 300 (30 X 10)	-.06	.73	-.02	.85	-.02	.79	.01	.87	.56	63.29	-1.26	39.48
	N = 900 (30 X 30)	.18	1.92	-.09	6.41	.10	2.00	-.14	6.18	.18	4.78	.24	12.45
	N = 1500 (30 X 50)	-.01	.39	.00	.32	.01	.35	.00	.31	-.01	.37	-.01	.33
	N = 500 (50 X 10)	-.08	2.08	.03	.83	-.01	1.82	-.06	.83	.02	20.43	.03	3.76
	N = 1500 (50 X 30)	.02	1.07	.01	.52	-.02	1.05	.03	.48	.03	.98	.02	.56
	N = 2500 (50 X 50)	-.01	.99	-.01	1.45	-.03	1.00	-.05	1.44	.00	1.01	.03	1.55
	N = 1000 (100 X 10)	-.09	2.32	.01	.53	.05	2.04	.01	.50	.01	2.36	.00	.54
	N = 3000 (100 X 30)	4.43	110.52	-.02	5.05	.69	104.67	-.10	5.00	-2.90	109.83	-.20	5.25
	N = 5000 (100 X 50)	-.01	.43	-.04	.55	-.02	.41	-.02	.55	-.01	.43	.01	.60
Moderately Nonnormal	N = 300 (30 X 10)	-.11	.87	.12	1.45	-.11	.91	.12	1.43	.03	14.04	.11	41.11
	N = 900 (30 X 30)	-.22	.84	1.10	9.56	-.19	.84	1.12	10.38	-.18	1.99	1.98	64.45
	N = 1500 (30 X 50)	-.02	.33	.00	.21	-.01	.34	.00	.23	-.01	.35	.00	.36
	N = 500 (50 X 10)	-.01	2.98	.06	1.34	-.22	4.14	.03	1.40	-.40	75.87	.62	38.14
	N = 1500 (50 X 30)	.60	2.10	.18	1.08	.55	2.11	.17	1.10	.60	2.27	.17	1.21
	N = 2500 (50 X 50)	.12	1.11	.26	2.33	.11	1.11	.31	2.42	.11	1.10	.35	2.46
	N = 1000 (100 X 10)	-.78	3.86	.07	.63	-.89	4.12	.07	.68	-.84	4.12	.05	.91
	N = 3000 (100 X 30)	27.55	165.72	-.77	9.26	-27.81	156.26	-.95	8.45	-27.13	157.84	-.97	8.55
	N = 5000 (100 X 50)	.02	.53	-.02	.67	.02	.53	.00	.66	.01	.53	-.04	.69
Severely Nonnormal	N = 300 (30 X 10)	-.06	.63	-.09	1.45	-.04	.80	-.11	1.61	.54	17.62	2.36	121.12
	N = 900 (30 X 30)	-.57	2.54	-.14	7.22	-.50	2.56	-.23	7.70	-.73	5.24	-.60	11.17
	N = 1500 (30 X 50)	.03	.51	.07	.30	.03	.51	.06	.29	.03	.58	.06	.41
	N = 500 (50 X 10)	-.04	1.49	-.09	1.33	.01	1.47	-.08	1.31	-.03	6.29	-.17	17.42
	N = 1500 (50 X 30)	-.15	2.41	-.19	.84	-.15	2.39	-.18	.84	-.12	2.51	-.21	.86
	N = 2500 (50 X 50)	-.13	1.41	-.18	2.33	-.14	1.42	-.23	2.35	-.17	1.50	-.24	2.39
	N = 1000 (100 X 10)	-.10	1.60	.02	.52	-.08	1.61	.02	.51	-.08	1.63	.02	.54
	N = 3000 (100 X 30)	27.85	165.80	-.06	10.01	-30.05	169.81	-.09	10.06	-31.09	169.14	-.04	10.16
	N = 5000 (100 X 50)	-.06	.59	-.11	.92	-.05	.59	-.10	.93	-.04	.58	-.11	1.00

Tables 32 through 34 display the descriptive results of bias among the standard errors with varying data normality conditions, sample size, and estimation methods. Regarding the level-1 standard errors, bias was similar across estimation methods with the largest standard deviation reported for the weighted least squares estimator in small sample size. Further, the weighted least squares estimator produced slightly less biased standard errors in moderate sample sizes, especially when cluster size ranged from 30 to 50 in severely non-normal data. Similar results were reported among level-2 standard errors.

The descriptive results examining the interaction standard errors among data normality conditions, sample size, and estimation method revealed that in small sample sizes, bias varied considerably by estimation method, with the most inconsistent results noted for the weighted least squares estimator. When the data increased in departure from normality, the generalized and weighted least squares estimators produced more biased results in small to moderate sample sizes.

As sample size increased in normal to moderately nonnormal conditions, the weighted least squares estimator produced slightly less biased standard errors. Under severe nonnormality conditions, the generalized and weighted least squares estimator appeared to produce more biased standard errors when compared to the maximum likelihood estimator. When the number of clusters was held at 100, and the cluster size ranged from 10 to 30, the weighted least squares estimator produced slightly less biased standard errors when compared to both the maximum likelihood and generalized least squares estimators. However, when the sample size increased to 5,000, the maximum likelihood and generalized least squares estimators produced slightly less biased standard errors than did the weighted least squares estimator. Note the largest standard deviation in each condition examined was associated with the weighted least squares estimator.

Table 32

Descriptive Results of Bias among Level-1 Standard Errors and Selected Design Conditions

		Estimation Method											
		ML				GLS				WLS			
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	-.05	.54	-.05	.54	.00	.56	.00	.56	-.10	.63	-.09	.61
	N = 900 (30 x 30)	.32	.22	.32	.22	.36	.29	.35	.30	.33	.32	.33	.33
	N = 1500 (30 x 50)	.00	.34	.00	.33	.01	.34	.01	.34	-.01	.35	.00	.35
	N = 500 (50 x 10)	-.17	.64	-.17	.64	-.14	.65	-.14	.65	-.19	.73	-.16	.68
	N = 1500 (50 x 30)	-.10	.56	-.10	.56	-.10	.56	-.10	.56	-.11	.56	-.11	.56
	N = 2500 (50 x 50)	.12	.30	.12	.30	.12	.30	.12	.30	.11	.30	.11	.30
	N = 1000 (100 x 10)	.15	.29	.15	.29	.16	.29	.16	.29	.13	.30	.13	.30
	N =3000 (100 x 30)	-.31	.43	-.31	.43	-.31	.43	-.31	.43	-.32	.43	-.33	.44
	N = 5000 (100 x 50)	-.03	.12	-2.79	3.83	-.03	.12	-2.79	3.83	-.03	.12	-2.81	3.86
Moderately Nonormal	N = 300 (30 x 10)	-.04	.63	-.04	.63	.06	.66	.06	.67	-.04	.78	-.04	.79
	N = 900 (30 x 30)	-.16	.60	-.16	.60	-.15	.61	-.15	.61	-.17	.61	-.16	.62
	N = 1500 (30 x 50)	-.34	.71	-.34	.71	-.33	.71	-.33	.71	-.31	.71	-.32	.71
	N = 500 (50 x 10)	-.62	1.78	-.62	1.78	-.54	1.73	-.54	1.73	-.57	1.84	-.55	1.78
	N = 1500 (50 x 30)	.16	.29	.16	.29	.17	.29	.17	.29	.16	.29	.17	.29
	N = 2500 (50 x 50)	.11	.22	.11	.22	.11	.23	.11	.22	.12	.22	.12	.22
	N = 1000 (100 x 10)	-.03	.48	-.03	.48	-.02	.49	-.02	.49	-.03	.47	-.03	.47
	N =3000 (100 x 30)	-.31	.42	-.32	.42	-.31	.41	-.31	.41	-.31	.41	-.30	.40
	N = 5000 (100 x 50)	.06	.18	-2.78	4.06	.06	.18	-2.78	4.06	.08	.18	-2.74	4.02
Severely Nonnormal	N = 300 (30 x 10)	-.45	1.15	-.45	1.15	-.41	1.18	-.41	1.18	-.47	1.19	-.48	1.19
	N = 900 (30 x 30)	-.21	.69	-.21	.69	-.26	1.06	-.26	1.06	-.19	.77	-.17	.72
	N = 1500 (30 x 50)	-.22	.67	-.22	.67	-.21	.68	-.21	.68	-.11	.60	-.08	.55
	N = 500 (50 x 10)	-.21	.79	-.21	.79	-.17	.80	-.18	.80	-.15	.73	-.16	.77
	N = 1500 (50 x 30)	.19	.19	.19	.19	.20	.20	.20	.20	.24	.19	.25	.18
	N = 2500 (50 x 50)	.14	.22	.14	.22	.14	.22	.14	.22	.21	.21	.19	.22
	N = 1000 (100 x 10)	-.30	.85	-.30	.85	-.29	.84	-.29	.85	-.26	.78	-.26	.77
	N =3000 (100 x 30)	-.33	.25	-.33	.25	-.33	.25	-.33	.25	-.24	.24	-.24	.22
	N = 5000 (100 x 50)	.04	.17	-2.83	4.09	.04	.17	-2.82	4.09	.13	.14	-2.41	3.61

Table 33

Descriptive Results of Bias among Level-2 Standard Errors and Selected Design Conditions

		Estimation Method											
		ML				GLS				WLS			
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Normal	N = 300 (30 X 10)	.36	2.08	.32	1.87	.28	1.96	.24	1.76	.23	1.91	.19	1.71
	N = 900 (30 X 30)	-.84	.12	-.71	.13	-.84	.12	-.72	.13	-.84	.12	-.72	.13
	N = 1500 (30 X 50)	-.40	.51	-.22	.44	-.39	.54	-.22	.44	-.39	.53	-.22	.45
	N = 500 (50 X 10)	-.31	.55	-.26	.49	-.32	.55	-.27	.49	-.33	.54	-.28	.48
	N = 1500 (50 X 30)	-.09	.97	-.03	.72	-.10	.96	-.03	.71	-.08	1.00	-.01	.75
	N = 2500 (50 X 50)	.49	1.47	.19	.74	.48	1.43	.18	.72	.49	1.47	.19	.74
	N = 1000 (100 X 10)	.14	.91	.20	.66	.15	.91	.21	.67	.13	.91	.19	.66
	N = 3000 (100 X 30)	.92	1.19	.65	.84	.91	1.18	.64	.83	.91	1.18	.64	.83
	N = 5000 (100 X 50)	-.20	.32	-.08	.15	-.20	.32	-.08	.15	-.20	.32	-.08	.15
Moderately Nonnormal	N = 300 (30 X 10)	-.52	1.23	-.48	1.18	-.50	1.26	-.47	1.20	-.53	1.24	-.50	1.19
	N = 900 (30 X 30)	-.53	.70	-.48	.65	-.53	.70	-.48	.66	-.54	.69	-.49	.65
	N = 1500 (30 X 50)	.18	1.41	-.03	.82	.18	1.41	-.03	.82	.16	1.41	-.04	.82
	N = 500 (50 X 10)	-.10	1.02	-.06	.87	-.06	1.06	-.03	.91	-.11	.99	-.09	.84
	N = 1500 (50 X 30)	-.06	.61	.04	.43	-.05	.61	.05	.43	-.08	.59	.03	.42
	N = 2500 (50 X 50)	.00	.88	.00	.56	.01	.90	.01	.57	.00	.88	.00	.56
	N = 1000 (100 X 10)	.05	.79	.12	.57	.05	.79	.12	.57	.03	.79	.10	.57
	N = 3000 (100 X 30)	.24	.64	.27	.43	.23	.64	.26	.42	.23	.64	.26	.43
	N = 5000 (100 X 50)	-.02	.65	-.02	.27	-.02	.66	-.01	.28	-.02	.66	-.02	.27
Severely Nonnormal	N = 300 (30 X 10)	-.21	1.50	-.18	1.42	-.11	1.80	-.09	1.70	-.18	1.73	-.16	1.63
	N = 900 (30 X 30)	-.73	.28	-.66	.26	-.74	.27	-.66	.25	-.75	.27	-.68	.25
	N = 1500 (30 X 50)	-.27	.65	-.23	.51	-.28	.65	-.23	.50	-.28	.65	-.24	.51
	N = 500 (50 X 10)	.08	1.65	.10	1.39	.08	1.68	.09	1.42	.03	1.62	.04	1.36
	N = 1500 (50 X 30)	.14	.90	.06	.56	.13	.89	.06	.56	.11	.88	.04	.55
	N = 2500 (50 X 50)	.10	1.38	-.04	.71	.10	1.37	-.04	.71	.08	1.35	-.05	.70
	N = 1000 (100 X 10)	-.10	.75	-.02	.58	-.09	.76	-.02	.59	-.13	.73	-.05	.57
	N = 3000 (100 X 30)	.12	.56	.13	.26	.11	.55	.13	.26	.09	.55	.11	.26
	N = 5000 (100 X 50)	-.43	.26	-.22	.19	-.43	.26	-.22	.18	-.44	.25	-.23	.19

Table 34

Descriptive Results of Bias among Cross-Level Interaction Standard Errors and Selected Design Conditions

		Estimation Method											
		ML				GLS				WLS			
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Normal	N = 300 (30 X 10)	.22	1.35	.14	1.07	.96	7.82	.76	6.17	2.38	45.43	1.81	34.63
	N = 900 (30 X 30)	.31	.46	1.04	1.23	.44	2.69	1.17	3.26	.37	2.35	1.10	2.99
	N = 1500 (30 X 50)	.23	.36	.47	1.07	.24	.38	.49	1.12	.20	.39	.46	1.10
	N = 500 (50 X 10)	.41	1.83	.30	1.36	.52	2.07	.41	1.79	.64	4.43	.52	3.64
	N = 1500 (50 X 30)	.13	.46	.25	.53	.13	.46	.25	.53	.09	.45	.21	.52
	N = 2500 (50 X 50)	.26	.34	-.13	.26	.26	.34	-.13	.26	.24	.33	-.15	.25
	N = 1000 (100 X 10)	.34	.50	.37	.33	.37	.53	.39	.34	.30	.51	.32	.34
	N = 3000 (100 X 30)	.04	.29	-.25	.19	.04	.29	-.25	.18	.03	.29	-.26	.18
	N = 5000 (100 X 50)	-.02	.11	.00	.29	-.02	.11	.01	.30	-.03	.11	-.01	.29
Moderately Nonnormal	N = 300 (30 X 10)	.44	2.66	.50	1.96	2.02	19.50	1.96	16.30	1.89	23.73	2.14	29.38
	N = 900 (30 X 30)	.01	.42	.65	1.48	.35	9.39	2.03	39.65	.59	19.49	5.22	124.23
	N = 1500 (30 X 50)	.04	.42	-.08	.63	.04	.42	-.06	.70	.03	.44	-.06	.89
	N = 500 (50 X 10)	.22	1.30	.18	.98	1.82	51.62	1.28	35.50	.52	5.49	.45	7.26
	N = 1500 (50 X 30)	.37	.56	.25	.55	.38	.61	.26	.57	.36	.64	.24	.57
	N = 2500 (50 X 50)	.20	.28	.19	.66	.20	.28	.19	.66	.20	.28	.19	.67
	N = 1000 (100 X 10)	.16	.46	.28	.71	.24	2.79	.37	3.85	.15	.70	.27	.96
	N = 3000 (100 X 30)	.01	.26	-.06	.40	.01	.26	-.06	.40	.00	.26	-.06	.40
	N = 5000 (100 X 50)	.10	.21	.09	.48	.10	.21	.09	.48	.11	.22	.10	.48
Severely Nonnormal	N = 300 (30 X 10)	.08	.86	.19	.86	.71	6.79	.73	5.55	1.20	30.70	1.13	24.37
	N = 900 (30 X 30)	.24	.55	.52	.90	.32	1.76	.58	1.76	.25	1.78	.51	1.79
	N = 1500 (30 X 50)	.10	.38	-.06	.52	.13	.57	-.03	.63	.19	.58	.05	.70
	N = 500 (50 X 10)	.13	.59	.13	.66	1.09	37.44	.97	30.15	.50	13.11	.59	16.08
	N = 1500 (50 X 30)	.35	.36	.13	.46	.38	.94	.15	.61	.38	.78	.17	.57
	N = 2500 (50 X 50)	.25	.27	.08	.55	.26	.36	.09	.59	.30	.30	.15	.63
	N = 1000 (100 X 10)	.02	.43	.13	.50	.03	.43	.14	.51	-.03	.40	.09	.48
	N = 3000 (100 X 30)	-.03	.20	-.04	.44	-.03	.21	-.05	.43	.00	.22	-.01	.46
	N = 5000 (100 X 50)	.04	.17	.35	.63	.04	.17	.35	.63	.11	.18	.48	.72

Bivariate Results

The Pearson product–moment correlations between parameter estimates, standard errors, and design conditions including the number of clusters, cluster size, correlation among variables, and data normality conditions are displayed in Table 35. Statistically significant correlations between the initial level-1 parameter estimate bias and design conditions ranged from .011, $p < .01$, with kurtosis (95% CI ranged from .006 to .015), to .029, $p < .01$, (95%CI ranged from .024 to .034), with skew. Similar results were found for the subsequent level-1 parameter estimate. Regarding the level-2 parameter estimates, none of the design conditions examined shared a statistically significant relation. Related to the cross-level interaction term, statistically significant associations ranged from -.048, $p <.01$, between the initial cross-level interaction terms and skew (95% CI ranged from -.053 to .043), to -.001 between the subsequent cross-level interaction term and correlation among variables (95% CI ranged from -.005 to .004).

Note that the effect size or percent of variance explained among each of the statistically significant correlations related to parameter estimate bias and design conditions was less than one percent.

Similar results were found when comparing level-1 parameter estimates and level-1 standard errors. Namely, the correlation among the level-1 standard errors and the design conditions were small, ranging from -.021, $p < .01$ between the subsequent level-1 standard error and cluster size (95% CI ranged from -.025 to –

016) to $r = -.004$, $p < .01$ between both level-1 standard errors and skew (95% CI ranged from -.008 to .008).

Statistically significant correlations between the level-2 standard errors and the study's design conditions ranged from .057 between skew and the subsequent level-2 parameter estimate’s standard error (95% CI ranged from .052 to.062) to .152 between the number of clusters, and the initial level-2 standard error 95% CI ranged from .149 to .152). Similar results were found for the subsequent level-2 parameter estimate’s standard with cluster size sharing the strongest association with both the initial and subsequent level-2 standard errors.

Although slight in magnitude, statistically significant results among the conditions examined and the standard errors associated with the cross-level interaction terms ranged from .019 between the initial cross-level interaction term and the number of clusters to -.006 between the subsequent cross-level interaction terms and the correlation among variables. Similar to the level-1 parameter estimates, the results should be interpreted with caution as each statistically significant correlation was associated with an effect size of less than one percent.

Table 35

Pearson Product-Moment Correlations among Parameter Estimate and Standard Error Bias and Selected Design Conditions

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Bias Level - 1 Predictor (1)	1.00																
Bias Level - 1 Predictor (2)	-.168**	1.00															
Bias Level - 2 Predictor (3)	-.007**	.002	1.00														
Bias Level - 2 Predictor (4)	.006**	.002	-.558**	1.00													
Bias Cross-Level Interact. (5)	-.068**	.000	.008**	.002	1.00												
Bias Cross-Level Interact. (6)	.477**	-.109**	.003	-.004*	.073**	1.00											
Bias Level - 1 Predictor SE (7)	-.038**	.053**	.001	-.001	.028**	.022**	1.00										
Bias Level - 1 Predictor SE (8)	-.040**	.053**	.001	-.001	.028**	.018**	1.000**	1.00									
Bias Level - 2 Predictor SE (9)	.007**	.065**	.013**	-.005*	-.008**	-.003	.004	.004	1.00								
Bias Level - 2 Predictor SE (10)	-.002	.077**	.015**	-.007**	-.003	-.003	.004*	.004	.952**	1.00							
Bias Cross-Level Interact. SE (11	-.038**	.023**	.003	-.004	.047**	-.022**	.862**	.869**	.018**	.020**	1.00						
Bias Cross-Level Interact. SE (12	-.001	.014**	-.001	-.001	.032**	.102**	.898**	.891**	-.009**	-.003	.694**	1.00					
No. of Clusters (13)	.000	.111**	-.002	.000	-.071**	-.006**	-.017**	-.020**	.103**	.152**	-.019**	-.017**	1.00				
Cluster Size (14)	.020**	-.140**	-.002	.000	-.003	-.003	-.019**	-.021**	-.012**	-.007**	-.020**	-.015**	.075**	1.00			
Correlation among Variables (15)	.018**	.002	.004	-.001	-.004	-.001	.002	.005*	-.054**	-.040**	.003	-.006**	.042**	.049**	1.00		
Skew (16)	.029**	-.016**	.000	.001	-.048**	-.004	-.004	-.004	-.045**	-.057**	-.006**	-.007**	-.001	-.001	.003	1.00	
Kurtosis (17)	.011**	-.090**	.001	-.001	.009**	.003	-.007**	-.007**	-.004	-.004	-.006**	-.008**	.001	.000	.001	.404**	1.00

**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed). The sample size for each coefficient was 181,875.

Factorial Analysis Results of Parameter Estimates

The factorial analysis results of the six study conditions and their effect on bias among the level-1 parameter estimates in model-three are displayed in Tables 36 and 37. The results revealed that estimation method as a main effect and the interaction effects that included estimation method by the study's design conditions were not statistically significant factors explaining variation in bias among the initial level–1 parameter estimate. In contrast, the remaining conditions were statistically significant, but explained less than one percent of the variation in the outcome variable. In regard to the subsequent level–1 parameter estimate, the estimation method was not statistically significant when entered as a main effect or interaction term with the remaining design conditions.

The correlation among variables explained the largest amount of variation in bias among the subsequent level−1 parameter estimate (partial $\eta^2 = .043$), while cluster size and the number of clusters explained approximately three percent of the variance in the outcome variable when entered as a main effect (partial $\eta^2 = .03$). Results of the interaction effects indicated that the number of clusters by correlation among variables explained approximately 11% of the variance in the criterion variable, while the number of clusters by cluster size and the number of clusters by correlation among variables were responsible for approximately nine percent of the variance in the subsequent level-1 parameter estimate. To gain further insight, the results for the subsequent level−1 parameter estimate, confidence intervals were calculated and are displayed in Figure 9.

Table 36
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Initial Level-1 Parameter Estimate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	471763.702	88	5361.0	74.49	< .001	.033[a]
Intercept	7710.149	1	7710.1	107.1	< .001	.001
Est. Mthd	14.539	2	7.270	.101	.904	.000
Clstrs	33953.505	2	16977	235.9	< .001	.002
Clstr. Size	12668.610	2	6334.3	88.01	< .001	.001
Corr	5835.336	2	2917.7	40.54	< .001	.000
Skew	10737.976	3	3579.3	49.73	< .001	.001
Kurt	11819.033	3	3939.7	54.74	< .001	.001
Est. Mthd. * Clstrs	62.590	4	15.647	.217	.929	.000
Est. Mthd. * Clstr. Size	70.758	4	17.690	.246	.912	.000
Est. Mthd. * Corr	129.967	4	32.492	.451	.771	.000
Est. Mthd. * Skew	450.042	6	75.007	1.042	.395	.000
Est. Mthd. * Kurt	207.431	6	34.572	.480	.823	.000
Clstrs * Clstr. Size	130197.688	4	32549	452.3	< .001	.009
Clstrs * Corr	56278.796	4	14070	195.5	< .001	.004
Clstrs * Skew	28410.204	6	4735.0	65.79	< .001	.002
Clstrs * Kurt	30851.474	6	5141.9	71.45	< .001	.002
Clstr. Size * Corr	46629.841	4	11657	162.0	< .001	.003
Clstr. Size * Skew	22446.548	6	3741.1	51.98	< .001	.002
Clstr. Size * Kurt	27983.316	6	4663.9	64.80	< .001	.002
Corr * Skew	21652.859	6	3608.8	50.14	< .001	.002
Corr * Kurt	7062.659	6	1177.1	16.36	< .001	.001
Skew * Kurt	18889.593	2	9444.8	131.2	< .001	.001
Error	13999293.3	194516	71.970			
Total	14477690.4	194605				
Corrected Total	14471057.0	194604				

a. R Squared = .033 (Adjusted R Squared = .032)

Note. The model eta squared for the full factorial 6-way model was .104 with df=696. The model eta squared for only the main and the two-way interaction effects was .033 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .071(i.e., .104-.033).

Table 37
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Subsequent Level-1 Parameter Estimate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	84601526	88	961381	1231	< .001	.358[a]
Intercept	2491796	1	2E+006	3190	< .001	.016
Est. Mthd	155.473	2	77.737	.100	.905	.000
Clstrs	4817577	2	2E+006	3084	< .001	.031
Clstr. Size	5222554	2	3E+006	3343	< .001	.033
Corr	6869867	2	3E+006	4398	< .001	.043
Skew	143235.8	3	47745	61.13	< .001	.001
Kurt	1151427	3	383809	491.4	< .001	.008
Est. Mthd. * Clstrs	1738.101	4	434.53	.556	.694	.000
Est. Mthd. * Clstr. Size	206.560	4	51.640	.066	.992	.000
Est. Mthd. * Corr	859.838	4	214.96	.275	.894	.000
Est. Mthd. * Skew	1618.419	6	269.74	.345	.913	.000
Est. Mthd. * Kurt	830.086	6	138.35	.177	.983	.000
Clstrs * Clstr. Size	15456541	4	4E+006	4947	< .001	.092
Clstrs * Corr	14706065	4	4E+006	4707	< .001	.088
Clstrs * Skew	633188.2	6	105531	135.1	< .001	.004
Clstrs * Kurt	2595604	6	432601	553.8	< .001	.017
Clstr. Size * Corr	20441793	4	5E+006	6543	< .001	.119
Clstr. Size * Skew	813636.4	6	135606	173.6	< .001	.005
Clstr. Size * Kurt	3542739	6	590457	755.9	< .001	.023
Corr * Skew	692560.4	6	115427	147.8	< .001	.005
Corr * Kurt	2428507	6	404751	518.2	< .001	.016
Skew * Kurt	117969.0	2	58985	75.51	< .001	.001
Error	1.5E+008	194516	781.10			
Total	2.4E+008	194605				
Corrected Total	2.4E+008	194604				

a. R Squared = .358 (Adjusted R Squared = .357)

Note. The model eta squared for the full factorial 6-way model was .700 with df=696. The model eta squared for only the main and the two-way interaction effects was .358 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four, five, and the six-way interaction effects was .342(i.e., .700-.358).

The results displayed in Figure 9 indicate that the subsequent level–1 parameter estimate was negatively biased in smaller sample sizes, especially when cluster size was 50 and the number of clusters was 30. Negative bias was most notable when the data departed from normality (kurtosis > 0) and the correlation among variables ranged from r = .50 to r = .70. As the number of clusters increased to 100, bias decreased considerably with similar results reported across clusters and cluster size.

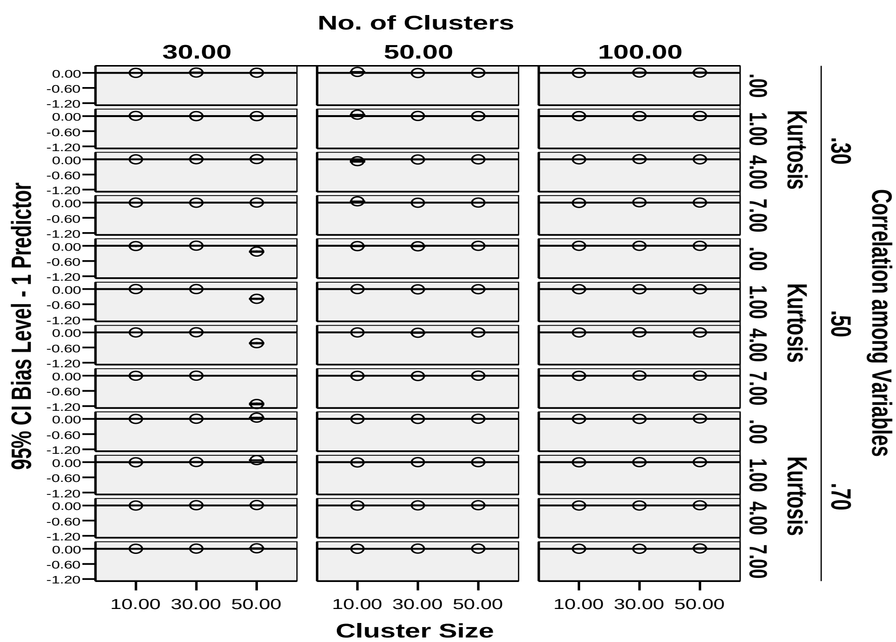

Figure 9. Confidence intervals examining bias among the subsequent level-1 parameter estimate in varying number of clusters, cluster size, kurtosis, and correlation among variables.

Tables 38 and 39 display the results examining bias among the level–2 parameter estimates. The results revealed that the estimation method and correlation among variables were statistically significant main effects, while none of the interaction effects had a statistically significant impact on the outcome variable. Although the estimation method and the correlation among variables were statistically significant, each explained less than one percent of the variation in the criterion variable. The resulting post hoc confidence intervals displayed in Figure 10 indicate that as the correlation among variables increased from r = .30 to r = .70, bias among the initial level-2 parameter estimate increased in variability under the weighted least squares estimator. Although bias was varied under the weighted least squares estimator, there was overlap among the confidence intervals.

Table 38
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Initial Level-2 Parameter Estimate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	28937685	88	328837	1.207	.091	.001[a]
Intercept	1102610	1	1E+006	4.046	.044	.000
Est. Mthd	1850515	2	925257	3.395	.034	.000
Clstrs	1188563	2	594281	2.180	.113	.000
Clstr. Size	1608181	2	804091	2.950	.052	.000
Corr	2146528	2	1E+006	3.938	.019	.000
Skew	434428.0	3	144809	.531	.661	.000
Kurt	770823.2	3	256941	.943	.419	.000
Est. Mthd. * Clstrs	1086132	4	271533	.996	.408	.000
Est. Mthd. * Clstr. Size	1356798	4	339199	1.245	.290	.000
Est. Mthd. * Corr	2564192	4	641048	2.352	.052	.000
Est. Mthd. * Skew	816337.0	6	136056	.499	.809	.000
Est. Mthd. * Kurt	1612261	6	268710	.986	.433	.000
Clstrs * Clstr. Size	1009545	4	252386	.926	.448	.000
Clstrs * Corr	815784.2	4	203946	.748	.559	.000
Clstrs * Skew	1465543	6	244257	.896	.496	.000
Clstrs * Kurt	1934232	6	322372	1.183	.312	.000
Clstr. Size * Corr	2002818	4	500704	1.837	.119	.000
Clstr. Size * Skew	1909646	6	318274	1.168	.320	.000
Clstr. Size * Kurt	2230147	6	371691	1.364	.225	.000
Corr * Skew	1708691	6	284782	1.045	.394	.000
Corr * Kurt	3009461	6	501577	1.840	.087	.000
Skew * Kurt	997087.1	2	498544	1.829	.161	.000
Error	5.2E+010	191468	272547			
Total	5.2E+010	191557				
Corrected Total	5.2E+010	191556				

a. R Squared = .001 (Adjusted R Squared =

Note. The model eta squared for the full factorial 6-way model was .004 with df=696. The model eta squared for only the main and the two-way interaction effects was .001 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .003(i.e., .004-.001).

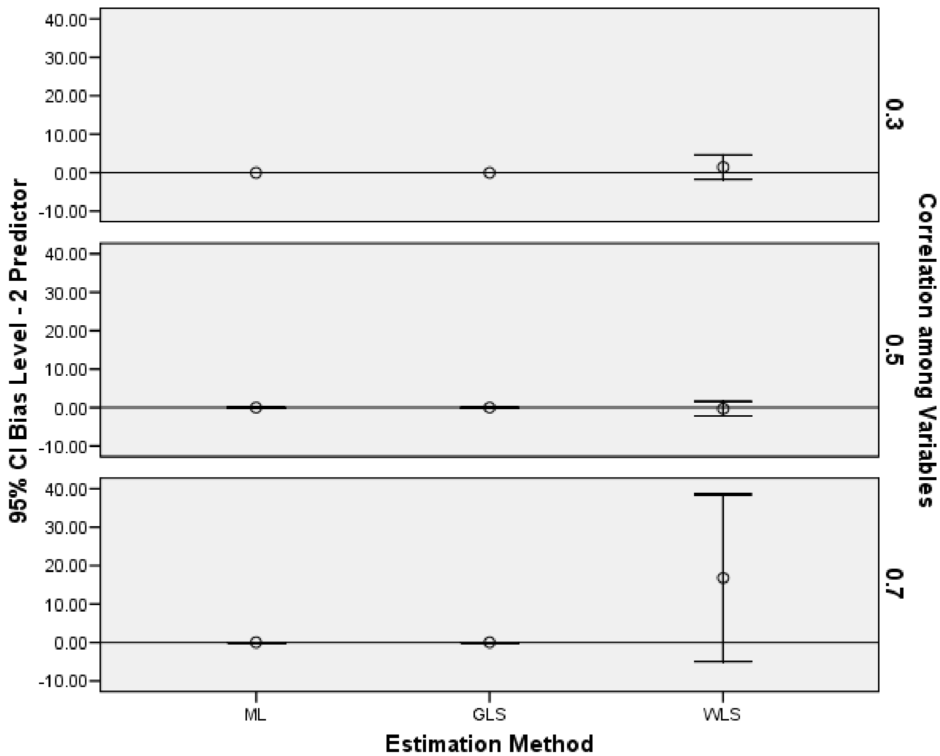

Figure 10. Confidence intervals examining bias among the initial level-2 parameter estimate by estimation method and correlation among variables.

The result for the subsequent level–2 parameter estimate in Table 36 revealed that none of the design conditions examined were statistically significant factors impacting bias. In more succinct terms, the results were similar to the results reported for both level–2 parameter estimates, as each design condition explained virtually no variation in the outcome variable.

Table 39
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Subsequent Level-2 Parameter Estimate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	5980847.4	88	67964	.126	.992	.000[a]
Intercept	31010.115	1	31010	.161	.688	.000
Est. Mthd	82031.171	2	41016	.213	.808	.000
Clstrs	197550.25	2	98775	.513	.599	.000
Clstr. Size	35350.926	2	17675	.092	.912	.000
Corr	14947.306	2	7473.7	.039	.962	.000
Skew	91277.314	3	30426	.158	.924	.000
Kurt	309647.29	3	103216	.536	.657	.000
Est. Mthd. * Clstrs	505578.13	4	126395	.657	.622	.000
Est. Mthd. * Clstr. Size	100390.45	4	25098	.130	.971	.000
Est. Mthd. * Corr	71132.913	4	17783	.092	.985	.000
Est. Mthd. * Skew	184589.92	6	30765	.160	.987	.000
Est. Mthd. * Kurt	695956.46	6	115993	.603	.728	.000
Clstrs * Clstr. Size	264582.29	4	66146	.344	.849	.000
Clstrs * Corr	346546.45	4	86637	.450	.772	.000
Clstrs * Skew	248936.56	6	41489	.216	.972	.000
Clstrs * Kurt	307664.78	6	51277	.266	.953	.000
Clstr. Size * Corr	144885.71	4	36221	.188	.945	.000
Clstr. Size * Skew	527417.27	6	87903	.457	.841	.000
Clstr. Size * Kurt	598003.83	6	99667	.518	.795	.000
Corr * Skew	144436.93	6	24073	.125	.993	.000
Corr * Kurt	949106.12	6	158184	.822	.553	.000
Skew * Kurt	243202.79	2	121601	.632	.532	.000
Error	3.72E+010	193338	192432			
Total	3.72E+010	193427				
Corrected Total	3.72E+010	193426				

a. R Squared = .000 (Adjusted R Squared = .000)

<u>Note</u>. The model eta squared for the full factorial 6-way model was .002 with <u>df</u>=696. The model eta squared for only the main and the two-way interaction effects was .000 with <u>df</u>=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .002(i.e., .002-.000).

Tables 40 and 41 displays the outcome of the design conditions on bias among the cross-level interaction parameter estimates. Regarding the interaction effect that was comprised of the initial level-1 and level-2 parameter estimates, the estimation method was not statistically significant as a main effect or when entered with the remaining design conditions in interaction terms. Although the remaining design conditions were statistically significant, each was responsible for less than one percent of the variation in the outcome variable. Related to the subsequent cross-level interaction parameter estimate, the number of clusters and kurtosis, when entered as main effects, were statistically significant, but explained less than one percent of variation in bias associated with the subsequent cross-level interaction parameter estimate. Similar to the initial cross-level interaction term, the statistically significant interaction term explained less than one percent of the variation in the dependent variable based on the resulting partial eta square.

Table 40
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Initial Interaction Parameter Estimate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	54059029	88	614307	238.8	< .001	.098[a]
Intercept	2953580	1	3E+006	1148	< .001	.006
Est. Mthd	55.708	2	27.854	.011	.989	.000
Clstrs	5414034	2	3E+006	1052	< .001	.011
Clstr. Size	5008419	2	3E+006	973.5	< .001	.010
Corr	5540340	2	3E+006	1077	< .001	.011
Skew	3200712	3	1E+006	414.7	< .001	.006
Kurt	827549.9	3	275850	107.2	< .001	.002
Est. Mthd. * Clstrs	5019.025	4	1254.8	.488	.745	.000
Est. Mthd. * Clstr. Size	6065.167	4	1516.3	.589	.670	.000
Est. Mthd. * Corr	3131.930	4	782.98	.304	.875	.000
Est. Mthd. * Skew	10525.60	6	1754.3	.682	.664	.000
Est. Mthd. * Kurt	12966.15	6	2161.0	.840	.539	.000
Clstrs * Clstr. Size	5818250	4	1E+006	565.4	< .001	.011
Clstrs * Corr	5610699	4	1E+006	545.3	< .001	.011
Clstrs * Skew	5908216	6	984703	382.8	< .001	.012
Clstrs * Kurt	1435985	6	239331	93.04	< .001	.003
Clstr. Size * Corr	5420388	4	1E+006	526.8	< .001	.011
Clstr. Size * Skew	5433433	6	905572	352.0	< .001	.011
Clstr. Size * Kurt	1504715	6	250786	97.49	< .001	.003
Corr * Skew	5253035	6	875506	340.3	< .001	.010
Corr * Kurt	1344993	6	224165	87.14	< .001	.003
Skew * Kurt	783311.0	2	391655	152.3	< .001	.002
Error	5.0E+008	194516	2572.4			
Total	5.6E+008	194605				
Corrected Total	5.5E+008	194604				

a. R Squared = .098 (Adjusted R Squared = .097)

Note. The model eta squared for the full factorial 6-way model was .303 with df=696. The model eta squared for only the main and the two-way interaction effects was .098 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .205(i.e., .303-.098).

Table 41
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Subsequent Cross-Level Interaction Parameter Estimate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	76323.30	88	867.31	1.695	< .001	.001[a]
Intercept	.005	1	.005	.000	.998	.000
Est. Mthd	218.424	2	109.21	.213	.808	.000
Clstrs	6735.487	2	3367.7	6.583	.001	.000
Clstr. Size	728.692	2	364.35	.712	.491	.000
Corr	206.650	2	103.32	.202	.817	.000
Skew	1480.348	3	493.45	.965	.408	.000
Kurt	6215.014	3	2071.7	4.049	.007	.000
Est. Mthd. * Clstrs	1184.059	4	296.01	.579	.678	.000
Est. Mthd. * Clstr. Size	1390.303	4	347.58	.679	.606	.000
Est. Mthd. * Corr	1235.481	4	308.87	.604	.660	.000
Est. Mthd. * Skew	1993.977	6	332.33	.650	.691	.000
Est. Mthd. * Kurt	3452.005	6	575.33	1.125	.345	.000
Clstrs * Clstr. Size	2113.911	4	528.48	1.033	.388	.000
Clstrs * Corr	3787.992	4	947.00	1.851	.116	.000
Clstrs * Skew	9417.006	6	1569.5	3.068	.005	.000
Clstrs * Kurt	1795.500	6	299.25	.585	.743	.000
Clstr. Size * Corr	3382.154	4	845.54	1.653	.158	.000
Clstr. Size * Skew	1761.317	6	293.55	.574	.752	.000
Clstr. Size * Kurt	2373.143	6	395.52	.773	.591	.000
Corr * Skew	5680.783	6	946.80	1.851	.085	.000
Corr * Kurt	6249.093	6	1041.5	2.036	.057	.000
Skew * Kurt	1650.764	2	825.38	1.613	.199	.000
Error	99512965	194516	511.59			
Total	99589520	194605				
Corrected Total	99589288	194604				

a. R Squared = .001 (Adjusted R Squared = .000)

<u>Note</u>. The model eta squared for the full factorial 6-way model was .004 with <u>df</u>=696. The model eta squared for only the main and the two-way interaction effects was .001 with <u>df</u>=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .003(i.e. ,.004-.001).

Factorial Analysis Results of Standard Errors

The results of the impact of the study's design conditions on the level-1 standard errors are displayed in Tables 42 and 43. The outcome revealed that the estimation method, the number of clusters and clusters size as main effects were statistically significant. In addition, statistically significant relations were reported for the interaction terms that included the estimation method by the number of clusters and estimation method by cluster size. Note that the interaction terms that included estimation method by skew and estimation method by kurtosis were not statistically significant. Each statistically significant effect explained less than one percent of the variation in bias among the level-1 parameter estimate standard errors. These results were expected based on earlier work by Hox and Maas (2001), who reported that bias was negligible among the level-1 parameter estimates and their associated standard errors. Similar results were found for both level-1 parameter estimate standard errors. Although statistically significant results were noted, post hoc confidence intervals were not calculated due to each statistically significant effect explaining virtually no variation in the outcome variable (partial η^2= 0.00).

Table 42
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Initial Level-1 Standard Error

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	148067.7	88	1682.6	5.998	< .001	.003[a]
Intercept	21377.39	1	21377	76.20	< .001	.000
Est. Mthd	4230.106	2	2115.1	7.539	< .001	.000
Clstrs	8145.006	2	4072.5	14.52	< .001	.000
Clstr. Size	13499.88	2	6749.9	24.06	< .001	.000
Corr	131.926	2	65.963	.235	.790	.000
Skew	1878.623	3	626.21	2.232	.082	.000
Kurt	1113.230	3	371.08	1.323	.265	.000
Est. Mthd. * Clstrs	6335.696	4	1583.9	5.646	< .001	.000
Est. Mthd. * Clstr. Size	6429.579	4	1607.4	5.730	< .001	.000
Est. Mthd. * Corr	1108.345	4	277.09	.988	.413	.000
Est. Mthd. * Skew	1196.019	6	199.34	.711	.641	.000
Est. Mthd. * Kurt	2170.283	6	361.71	1.289	.258	.000
Clstrs * Clstr. Size	7595.595	4	1898.9	6.769	< .001	.000
Clstrs * Corr	4210.098	4	1052.5	3.752	.005	.000
Clstrs * Skew	3533.629	6	588.94	2.099	.050	.000
Clstrs * Kurt	15650.20	6	2608.4	9.298	< .001	.000
Clstr. Size * Corr	4168.879	4	1042.2	3.715	.005	.000
Clstr. Size * Skew	3253.552	6	542.26	1.933	.072	.000
Clstr. Size * Kurt	4011.382	6	668.56	2.383	.026	.000
Corr * Skew	18343.88	6	3057.3	10.90	< .001	.000
Corr * Kurt	5675.611	6	945.94	3.372	.003	.000
Skew * Kurt	1049.905	2	524.95	1.871	.154	.000
Error	54569022	194516	280.54			
Total	54753447	194605				
Corrected Total	54717090	194604				

a. R Squared = .003 (Adjusted R Squared = .002)

Note. The model eta squared for the full factorial 6-way model was .011 with df=696. The model eta squared for only the main and the two-way interaction effects was .003 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .008(i.e. ,.011-.003).

Table 43
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Subsequent Level-1 Standard Error

Source	Type Sum Square	df	Mea Squar	F	Sig.	Partia Et Square
Corrected Model	151701.5	8	1723.	6.25	< .001	.003[a]
Intercept	17163.69	1	1716	62.3	< .001	.000
Est. Mthd	4183.63	2	2091.	7.59	< .001	.000
Clstrs	11467.60	2	5733.	20.8	< .001	.000
Clstr. Size	15955.97	2	7978.	28.9	< .001	.000
Corr	923.66	2	461.8	1.67	.187	.000
Skew	1763.45	3	587.8	2.13	.094	.000
Kurt	1032.38	3	344.1	1.24	.290	.000
Est. Mthd. * Clstrs	6246.92	4	1561.	5.67	< .001	.000
Est. Mthd. * Clstr. Size	6434.26	4	1608.	5.84	< .001	.000
Est. Mthd. * Corr	1125.86	4	281.4	1.02	.394	.000
Est. Mthd. * Skew	1157.58	6	192.9	.70	.649	.000
Est. Mthd. * Kurt	2121.03	6	353.5	1.28	.261	.000
Clstrs * Clstr. Size	3857.27	4	964.3	3.50	.007	.000
Clstrs * Corr	3263.28	4	815.8	2.96	.019	.000
Clstrs * Skew	3467.54	6	577.9	2.09	.050	.000
Clstrs * Kurt	15656.24	6	2609.	9.47	< .001	.000
Clstr. Size * Corr	3389.11	4	847.2	3.07	.015	.000
Clstr. Size * Skew	3111.97	6	518.6	1.88	.080	.000
Clstr. Size * Kurt	3779.98	6	630.0	2.28	.033	.000
Corr * Skew	17826.96	6	2971.	10.7	< .001	.000
Corr * Kurt	5451.74	6	908.6	3.29	.003	.000
Skew * Kurt	963.10	2	481.5	1.74	.174	.000
Error	5357645	19451	275.4			
Total	5375856	19460				
Corrected	5372815	19460				

[a]. R Squared = .003 (Adjusted R Squared =

Note. The model eta squared for the full factorial 6-way model was .011 with df=696. The model eta squared for only the main and the two-way interaction effects was .003 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .008(i.e. ,.011-.003).

Tables 44 and 45 illustrate the impact of the six design conditions on the level-2 parameter estimates' standard errors. Similar to the results reported for the two previous models, the number of clusters as a main effect had the greatest impact on bias among the level-2 parameter estimate standard errors (partial η^2= .099 and .120 respectively). The estimation method was statically significant as a main effect for both level-2 standard errors. When the estimation method was entered as an interaction term with the remaining design conditions, each was statistically significant for the initial level-2 standard error, while only the estimation method by the number of clusters and cluster size shared a statistically significant relation with the subsequent level-2 standard error. The correlation among variables as a main effect was statistically significant and explained approximately three percent of the variance in bias among the standard errors. Concerning interaction terms, the number of clusters by cluster size explained approximately five percent of the bias variance among both level-2 standard errors, suggesting that as the number of clusters and cluster size increased, the amount of bias in the outcome variable decreased.

To gain insight into the statistically significant results, post hoc confidence intervals were calculated. The results in Figures 11 and 12 revealed that as the number of clusters increased, the amount of bias decreased with the standard errors negatively biased in smaller sample sizes. In addition, as the correlation among variables increased from $r = .30$ to $r = .50$, the standard errors were negatively biased in smaller sample sizes. However, as the correlation among variables increased to $r = .70$, bias decreased in the smaller sample sizes. Further, when the number of clusters increased to 100, the standard errors were negatively biased when cluster size was held at 10. Similar results were noted for both level-2 parameter estimates' standard error.

Table 44
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Initial Level-2 Standard Error

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2649328.524	88	30106.006	913.789	< .001	.292[a]
Intercept	1221666.611	1	1221666.611	37080.472	< .001	.160
Est. Mthd	896.861	2	448.431	13.611	< .001	.000
Clstrs	707935.101	2	353967.551	10743.753	< .001	.099
Clstr. Size	39684.441	2	19842.221	602.258	< .001	.006
Corr	216790.008	2	108395.004	3290.045	< .001	.033
Skew	13427.621	3	4475.874	135.853	< .001	.002
Kurt	55547.053	3	18515.684	561.995	< .001	.009
Est. Mthd. * Clstrs	1461.924	4	365.481	11.093	< .001	.000
Est. Mthd. * Clstr. Size	719.784	4	179.946	5.462	< .001	.000
Est. Mthd. * Corr	340.691	4	85.173	2.585	.035	.000
Est. Mthd. * Skew	615.628	6	102.605	3.114	.005	.000
Est. Mthd. * Kurt	887.524	6	147.921	4.490	< .001	.000
Clstrs * Clstr. Size	318915.694	4	79728.924	2419.961	< .001	.047
Clstrs * Corr	322439.426	4	80609.857	2446.700	< .001	.048
Clstrs * Skew	10428.624	6	1738.104	52.756	< .001	.002
Clstrs * Kurt	60518.127	6	10086.355	306.145	< .001	.009
Clstr. Size * Corr	115593.022	4	28898.256	877.130	< .001	.018
Clstr. Size * Skew	97929.837	6	16321.640	495.400	< .001	.015
Clstr. Size * Kurt	126449.065	6	21074.844	639.671	< .001	.019
Corr * Skew	104015.456	6	17335.909	526.186	< .001	.016
Corr * Kurt	24245.504	6	4040.917	122.651	< .001	.004
Skew * Kurt	1088.578	2	544.289	16.520	< .001	.000
Error	6408594.339	194516	32.946			
Total	10711448.788	194605				
Corrected Total	9057922.864	194604				

a. R Squared = .292 (Adjusted R Squared = .292)

Note. The model eta squared for the full factorial 6-way model was .515 with df=696. The model eta squared for only the main and the two-way interaction effects was .292 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .223(i.e., .515-.292).

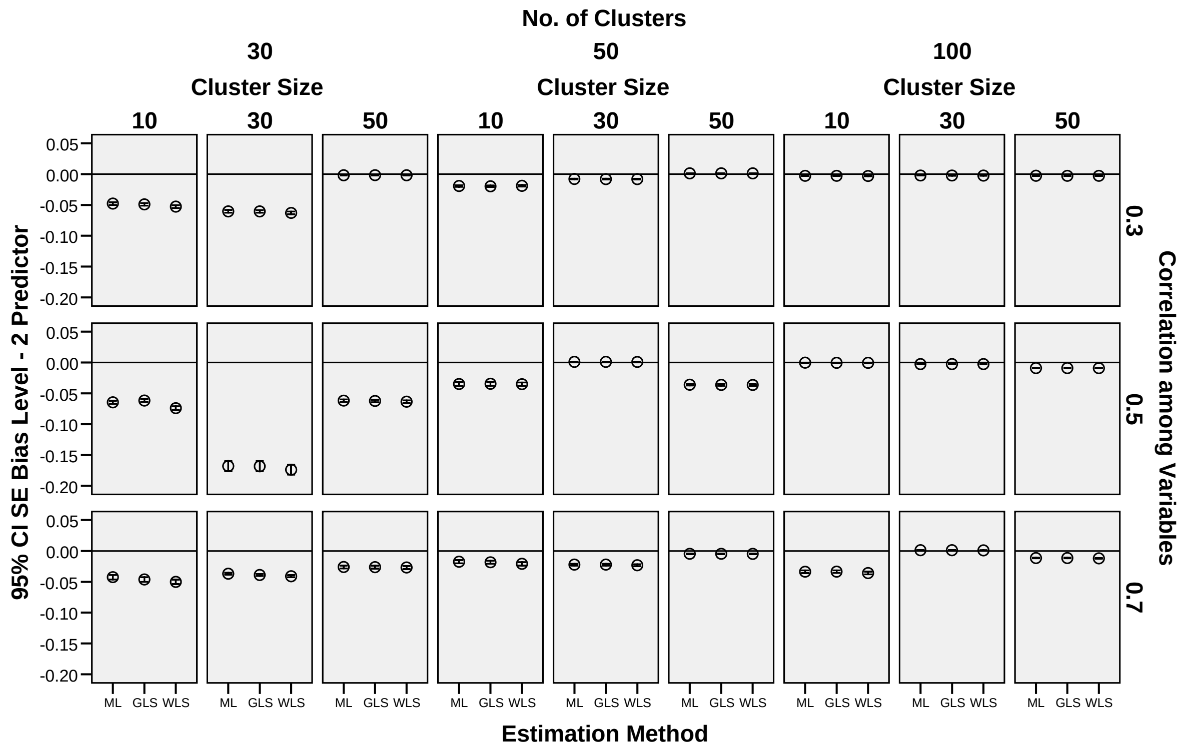

Figure 11. Confidence intervals examining bias among the initial level-1 standard error by estimation method, number of clusters, cluster size, and correlation among variables.

Table 45
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Subsequent Level-2 Standard Error

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	598636.862	88	6802.692	1197.938	< .001	.351[a]
Intercept	244212.575	1	244212.575	43005.248	< .001	.181
Est. Mthd	470.013	2	235.007	41.384	< .001	.000
Clstrs	152434.476	2	76217.238	13421.673	< .001	.121
Clstr. Size	23278.004	2	11639.002	2049.600	< .001	.021
Corr	36094.539	2	18047.270	3178.081	< .001	.032
Skew	1455.964	3	485.321	85.464	< .001	.001
Kurt	13507.169	3	4502.390	792.860	< .001	.012
Est. Mthd. * Clstrs	483.323	4	120.831	21.278	< .001	.000
Est. Mthd. * Clstr. Size	393.651	4	98.413	17.330	< .001	.000
Est. Mthd. * Corr	59.377	4	14.844	2.614	.033	.000
Est. Mthd. * Skew	29.853	6	4.975	.876	.511	.000
Est. Mthd. * Kurt	47.022	6	7.837	1.380	.218	.000
Clstrs * Clstr. Size	53492.867	4	13373.217	2354.991	< .001	.046
Clstrs * Corr	74797.193	4	18699.298	3292.902	< .001	.063
Clstrs * Skew	2812.132	6	468.689	82.535	< .001	.003
Clstrs * Kurt	15940.486	6	2656.748	467.847	< .001	.014
Clstr. Size * Corr	18851.481	4	4712.870	829.925	< .001	.017
Clstr. Size * Skew	9983.021	6	1663.837	292.998	< .001	.009
Clstr. Size * Kurt	34473.590	6	5745.598	1011.786	< .001	.030
Corr * Skew	15628.327	6	2604.721	458.685	< .001	.014
Corr * Kurt	3492.947	6	582.158	102.517	< .001	.003
Skew * Kurt	937.882	2	468.941	82.579	< .001	.001
Error	1104592.010	194516	5.679			
Total	2052008.017	194605				
Corrected Total	1703228.872	194604				

a. R Squared = .351 (Adjusted R Squared = .351)

Note. The model eta squared for the full factorial 6-way model was .590 with df=696. The model eta squared for only the main and the two-way interaction effects was .351 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .239(i.e., .590-.351).

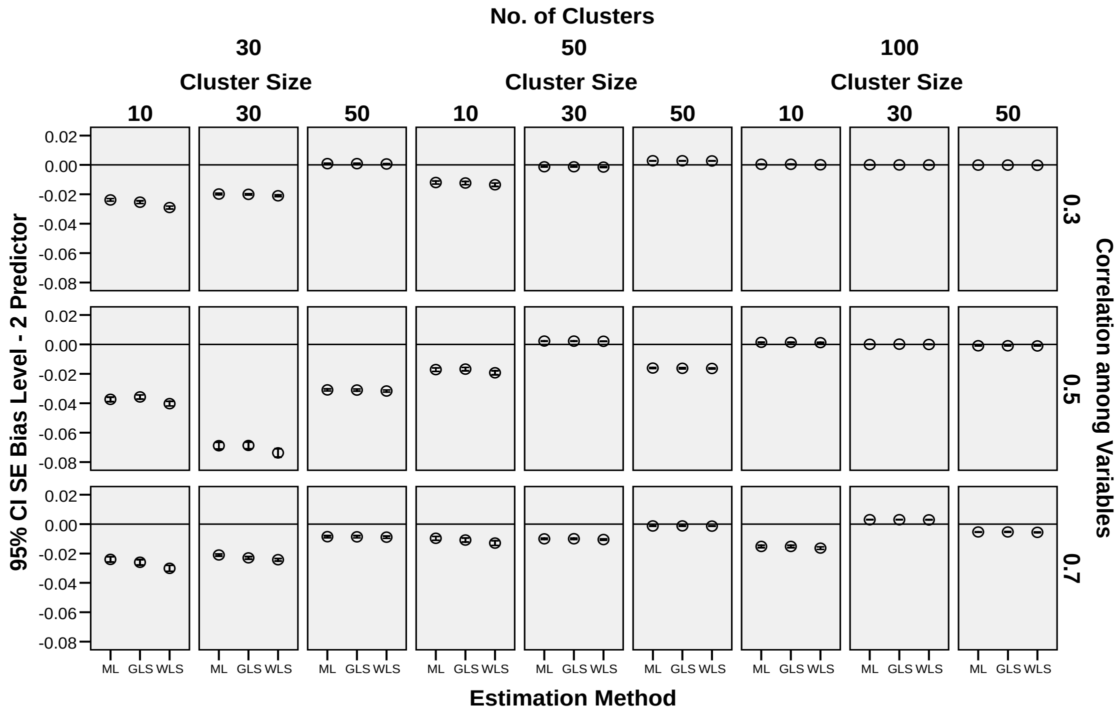

Figure 12. Confidence intervals examining bias among the subsequent level-1 standard error by estimation method, number of clusters, cluster size, and correlation among variables.

Tables 46 and 47 display the outcome of the cross-level interaction terms included the initial level-1 parameter estimate by the initial level-2 parameter estimate (Table 43), and the subsequent level-1 parameter estimates standard error by the subsequent level-2 parameter estimates standard error (Table 44). Regarding the initial cross-level standard error, with the exception of estimation method, each of the remaining design conditions was statistically significant, with cluster size explaining the largest amount of bias variance (approximately three percent [partial $\eta^2 = .024$]). As for interaction effects, each was statically significant with the term that included cluster size by correlation among variables possessing the largest partial eta square (partial $\eta^2 = .019$). Although each of the remaining interaction terms was statistically significant, each explained little variance in the outcome variable.

Similar effects were reported for the subsequent cross-level interaction standard errors in Table 44. One notable difference between the initial and subsequent cross-level interaction standard errors was that the estimation method was statistically significant as a main and interaction within effects for the subsequent standard error. However, similar to the results reported for models one and two, the estimation method explained a negligible portion of variance in the dependent variable. Relating to interaction effects, the number of clusters by cluster size was responsible for the greatest amount of bias variance in the subsequent cross-level interaction standard error (approximately three percent [partial $\eta^2 = .025$]).

Post hoc confidence intervals were derived to provide a pictorial view of the salient statistically significant effects. Figure 13 displays the confidence intervals related to the cluster size, estimation method, and the correlation among variables for the initial level-2 standard error. The results show that bias decreased as cluster size increased. Further, as the correlation among variables increased from r = .3 to r = .7, bias among the initial level-2 standard errors increased across cluster sizes, with a noted difference when cluster size was held at 10. Figure 14 exhibits the post hoc confidence intervals examining bias among the subsequent level-2 parameters estimate. The results were similar to those reported in Figure 13 (i.e., bias tended to decrease as the number of clusters increased). The most variation was noted in the cluster size of 30 and when the correlation among variables was held at r = .70. Note that the weighted least squares estimator produced more negatively biased standard errors in small clusters sizes.

Table 46
Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Initial Interaction Standard Error

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	23779.52	88	270.222	343.391	< .001	.134[a]
Intercept	1285.307	1	1285.31	1633.3	< .001	.008
Est. Mthd	2.553	2	1.276	1.622	.197	.000
Clstrs	323.008	2	161.504	205.235	< .001	.002
Clstr. Size	3836.894	2	1918.45	2437.9	< .001	.024
Corr	635.916	2	317.958	404.053	< .001	.004
Skew	88.400	3	29.467	37.446	< .001	.001
Kurt	1451.674	3	483.891	614.917	< .001	.009
Est. Mthd. * Clstrs	90.430	4	22.607	28.729	< .001	.001
Est. Mthd. * Clstr. Size	94.928	4	23.732	30.158	< .001	.001
Est. Mthd. * Corr	28.947	4	7.237	9.196	< .001	.000
Est. Mthd. * Skew	46.432	6	7.739	9.834	< .001	.000
Est. Mthd. * Kurt	31.839	6	5.307	6.743	< .001	.000
Clstrs * Clstr. Size	1140.806	4	285.201	362.427	< .001	.007
Clstrs * Corr	1405.436	4	351.359	446.498	< .001	.009
Clstrs * Skew	1430.309	6	238.385	302.933	< .001	.009
Clstrs * Kurt	1334.321	6	222.387	282.604	< .001	.009
Clstr. Size * Corr	2966.022	4	741.506	942.287	< .001	.019
Clstr. Size * Skew	198.510	6	33.085	42.044	< .001	.001
Clstr. Size * Kurt	1674.485	6	279.081	354.649	< .001	.011
Corr * Skew	2334.968	6	389.161	494.536	< .001	.015
Corr * Kurt	993.156	6	165.526	210.346	< .001	.006
Skew * Kurt	438.044	2	219.022	278.328	< .001	.003
Error	153068.8	194516	.787			
Total	178677.8	194605				
Corrected Total	176848.4	194604				

[a]. R Squared = .134 (Adjusted R Squared = .134)

Note. The model eta squared for the full factorial 6-way model was .403 with df=696. The model eta squared for only the main and the two-way interaction effects was .134 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .269(i.e., .403-.134).

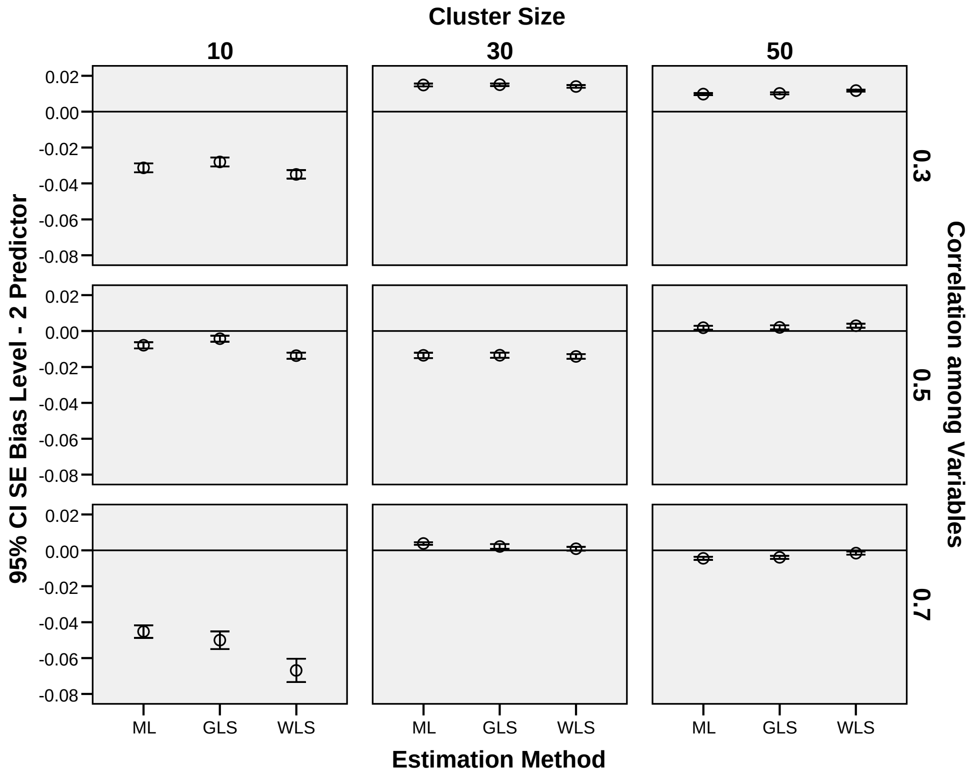

Figure 13. Confidence intervals examining bias among the initial level-2 standard error by estimation method, cluster size, and correlation among variables.

Table 47

Factorial ANOVA Results of the Study's Six Design Conditions Effect on Bias among the Subsequent Cross-Level Interaction Standard Error

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	26100.60	88	296.598	247.907	< .001	.101[a]
Intercept	5370.074	1	5370.07	4488.5	< .001	.023
Est. Mthd	.153	2	.076	.064	< .001	.000
Clstrs	340.159	2	170.079	142.158	< .001	.001
Clstr. Size	2181.155	2	1090.58	911.542	< .001	.009
Corr	1023.253	2	511.627	427.635	< .001	.004
Skew	63.286	3	21.095	17.632	< .001	.000
Kurt	680.994	3	226.998	189.733	< .001	.003
Est. Mthd. * Clstrs	90.383	4	22.596	18.886	< .001	.000
Est. Mthd. * Clstr. Size	112.662	4	28.166	23.542	< .001	.000
Est. Mthd. * Corr	42.410	4	10.602	8.862	< .001	.000
Est. Mthd. * Skew	53.043	6	8.841	7.389	< .001	.000
Est. Mthd. * Kurt	34.796	6	5.799	4.847	< .001	.000
Clstrs * Clstr. Size	5866.226	4	1466.56	1225.8	< .001	.025
Clstrs * Corr	1091.797	4	272.949	228.140	< .001	.005
Clstrs * Skew	884.580	6	147.430	123.227	< .001	.004
Clstrs * Kurt	1326.136	6	221.023	184.738	< .001	.006
Clstr. Size * Corr	2984.102	4	746.025	623.554	< .001	.013
Clstr. Size * Skew	532.905	6	88.818	74.237	< .001	.002
Clstr. Size * Kurt	1459.887	6	243.315	203.371	< .001	.006
Corr * Skew	3401.192	6	566.865	473.806	< .001	.014
Corr * Kurt	2117.545	6	352.924	294.986	< .001	.009
Skew * Kurt	337.372	2	168.686	140.993	< .001	.001
Error	232720.7	194516	1.196			
Total	265473.0	194605				
Corrected Total	258821.3	194604				

a. R Squared = .101 (Adjusted R Squared = .100)

Note. The model eta squared for the full factorial 6-way model was .364 with df=696. The model eta squared for only the main and the two-way interaction effects was .101 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four, five, and the six-way interaction effects was .206 (i.e., .364-.101).

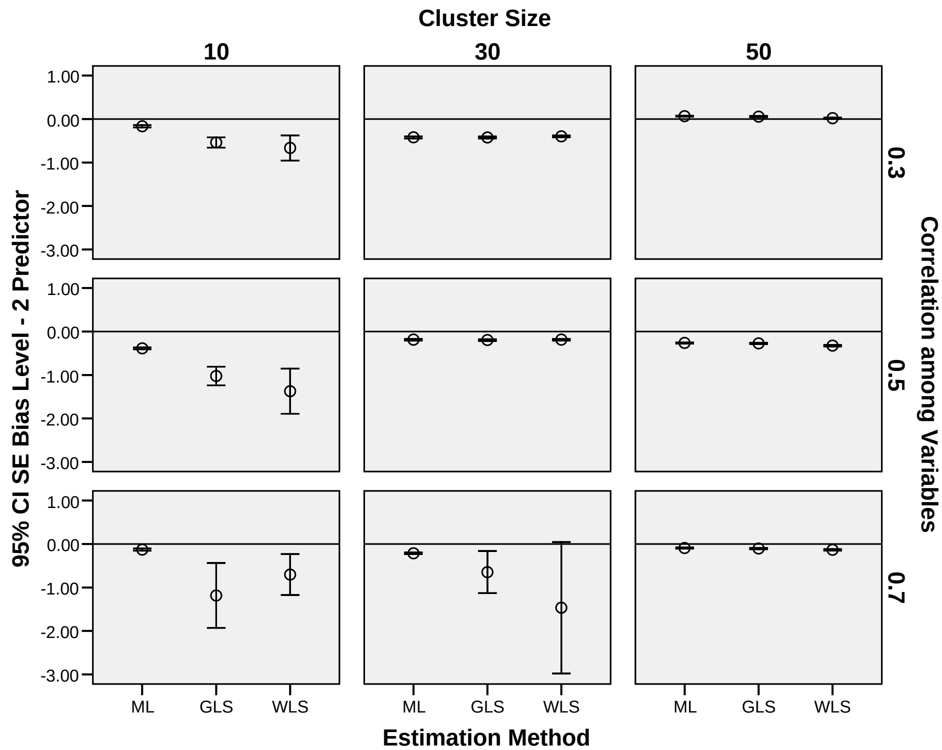

Figure 14. Confidence intervals examining bias among the subsequent interaction standard error by estimation method, cluster size, and correlation among variables.

Summary of Model Three

The results from model one indicated that of the six conditions examined, the number of clusters, cluster size and the correlation among variables had a statistically significant effect on bias among the level-1 parameter estimates and the cross-level interaction term, but explained little variation in the criterion variable. However, when investigating the standard errors, the number of clusters appeared to have the greatest impact on bias among the standard errors, especially among the level-2 standard errors. When examining interaction effects via confidence intervals as a post hoc procedure, bias was most prevalent when the number of clusters and cluster size were smaller. Most notably, when the number of clusters was 30 and 50 and cluster size ranged from 10 to 50, the standard errors were biased downward. In this same setting, when kurtosis was severe (kurtosis = 7), the weighted least square estimator produced slightly less biased results when compared to the maximum likelihood and generalized least squares estimators respectively. However, as sample size increased, bias decreased with each estimator producing similar results when the number of clusters increased to 100 regardless of cluster size.

Section II: Results of the Impact of the Six Study Design Conditions on Bias Associated with Selected Fit Indices

Section II includes the results of the descriptive, bivariate, and multivariate analyses of the six study design conditions' impact on bias among selected model fit indices. The selected model fit indices investigated included the goodness of fit index (GFI), root mean square residual (RMR), comparative fit index (CFI), fit, and the Akaike Information Criterion (AIC). The research question that guided this portion of the study was: Do fit indices differ by the design factors (estimation method, data normality conditions, and sample size)? The effect of the six study design conditions on fit index bias is examined in three models of varying complexity. The results are reported below.

Model One

Model one included one predictor variable at level-1 ($\gamma 1$), one predictor variable at level-2, and an cross-level interaction effect (level-1 x level-2). The model is displayed below.

Level-1 equation $Y_{ij} = \beta_0 + \beta_1 X1_{ij} + \varepsilon_{ij}$

Level-2 equation $\beta_0 = \gamma_{00} + U_{0j}$

$\beta_1 = \gamma_{10} + \gamma 11X\ W_1BAR_j$

Combined Equation $Y = \gamma_{00} + \gamma_0 X_1 + \gamma_{11} W_1BAR*X1 + U_0 + \varepsilon$

Descriptive Measures of Bias among Selected Fit Indices

Table 48 presents the descriptive measures of GIF bias by data normality conditions, sample size, and estimation method for model one. In normal data conditions and in small sample sizes, especially when the number of clusters equaled 30 and cluster size was 10, negative GFI bias was associated with each estimation method with the maximum likelihood and generalized least squares estimators producing similar results, while the weighted least squares estimator produced slightly more negatively biased results. As sample size increased, similar results were reported across estimation methods.

Mean GFI bias ranged from -.397 (*SD* = 4.766) for the weighted least squares estimator when the number of clusters equaled 30 and cluster size was 10 in severely nonnormal data to .001 (*SD* = .002) for the maximum likelihood estimator when the number of clusters and cluster size was held at 50 in normal data conditions.

Table 48

Descriptive Results of GFI Bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	-.042	.132	-.062	.262	-.104	.475
	N = 900 (30 x 30)	-.001	.017	-.001	.013	-.001	.013
	N = 1500 (30 x 50)	.003	.001	.002	.001	.003	.001
	N = 500 (50 x 10)	.005	.004	.005	.003	.005	.003
	N = 1500 (50 x 30)	.001	.002	.001	.003	.001	.003
	N = 2500 (50 x 50)	.003	.002	.003	.002	.004	.002
	N = 1000 (100 x 10)	.003	.003	.003	.003	.003	.003
	N = 3000 (100 x 30)	.002	.000	.002	.000	.002	.000
	N = 5000 (100 x 50)	.003	.001	.003	.001	.003	.001
Moderately Nonnormal	N = 300 (30 x 10)	-.016	.078	-.034	.247	-.038	.291
	N = 900 (30 x 30)	-.004	.043	-.003	.032	-.003	.035
	N = 1500 (30 x 50)	.002	.005	.002	.006	.002	.008
	N = 500 (50 x 10)	-.001	.041	.000	.029	.000	.028
	N = 1500 (50 x 30)	.001	.004	.001	.003	.001	.003
	N = 2500 (50 x 50)	.002	.001	.002	.001	.002	.001
	N = 1000 (100 x 10)	.002	.012	.003	.007	.003	.007
	N = 3000 (100 x 30)	.002	.000	.002	.000	.002	.000
	N = 5000 (100 x 50)	.003	.001	.003	.001	.003	.001
Severely Nonnormal	N = 300 (30 x 10)	-.019	.086	-.136	1.441	-.397	4.766
	N = 900 (30 x 30)	-.004	.056	-.013	.148	-.019	.203
	N = 1500 (30 x 50)	.002	.002	.002	.002	.002	.003
	N = 500 (50 x 10)	-.015	.092	-.031	.209	-.031	.235
	N = 1500 (50 x 30)	.002	.001	.002	.001	.001	.002
	N = 2500 (50 x 50)	.002	.001	.002	.001	.001	.002
	N = 1000 (100 x 10)	.002	.007	.002	.006	.001	.011
	N = 3000 (100 x 30)	.002	.001	.002	.001	.002	.001
	N = 5000 (100 x 50)	.003	.001	.003	.001	.003	.002

Regarding fit index bias reported in Table 49, the greatest amount of variation in fit index bias was in small sample sizes especially when the number of clusters equaled 30 and 50 while cluster size was held at 10 and 30 in each data normality condition investigated. As sample size increased, bias associated with the fit index decreased across estimation methods with similar results noted for the maximum likelihood and generalized least squares estimator. Although similar in sign, fit index bias was greater with the weighted least squares estimator in the majority of scenarios examined.

Fit bias ranged from -.676 (*SD* = .816) for the weighted least squares estimator when the number of clusters and cluster size was 50 in normal data conditions to 60.603 (*SD* = 674.384) for the weighted least squares estimator in severely nonnormal data conditions when the number of clusters was 30 and cluster size was held at 10.

Table 49

Descriptive Results of Fit Bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	24.397	148.112	14.929	65.301	24.111	105.542
	N = 900 (30 x 30)	.109	4.430	.093	3.206	.006	3.091
	N = 1500 (30 x 50)	-.593	.864	-.577	.802	-.768	.479
	N = 500 (50 x 10)	.092	4.180	.049	3.681	-.244	3.004
	N = 1500 (50 x 30)	.274	2.671	.500	3.365	.029	2.664
	N = 2500 (50 x 50)	-.426	1.173	-.376	1.352	-.637	.799
	N = 1000 (100 x 10)	-.614	.777	-.649	.701	-.676	.816
	N = 3000 (100 x 30)	-.410	.900	-.461	.784	-.564	.890
	N = 5000 (100 x 50)	.119	1.787	.174	1.848	-.269	1.201
Moderately Nonnormal	N = 300 (30 x 10)	13.196	254.369	7.933	61.539	8.985	73.732
	N = 900 (30 x 30)	3.270	36.442	1.794	14.577	1.223	11.296
	N = 1500 (30 x 50)	-.167	2.292	-.118	2.712	-.305	2.199
	N = 500 (50 x 10)	4.899	60.371	2.515	19.277	3.977	33.249
	N = 1500 (50 x 30)	1.306	9.098	1.130	7.855	.311	4.304
	N = 2500 (50 x 50)	-.345	1.571	-.378	1.338	-.588	.964
	N = 1000 (100 x 10)	-.335	6.156	-.472	2.813	-.533	2.312
	N = 3000 (100 x 30)	-.424	1.026	-.455	1.020	-.652	.820
	N = 5000 (100 x 50)	-.010	1.641	.059	1.916	-.378	1.286
Severely Nonnormal	N = 300 (30 x 10)	26.426	732.537	33.450	359.570	60.603	674.348
	N = 900 (30 x 30)	5.324	125.713	3.249	36.933	1.998	26.653
	N = 1500 (30 x 50)	.012	4.278	.035	4.238	-.219	5.380
	N = 500 (50 x 10)	18.896	298.730	12.318	86.361	6.622	38.914
	N = 1500 (50 x 30)	.026	1.975	.045	1.967	-.314	1.685
	N = 2500 (50 x 50)	-.316	1.645	-.300	1.650	-.589	1.197
	N = 1000 (100 x 10)	-.317	2.571	-.367	2.501	-.274	2.635
	N = 3000 (100 x 30)	-.352	1.839	-.381	2.168	-.642	.921
	N = 5000 (100 x 50)	.026	1.619	.056	1.711	-.405	1.354

The results reported in Table 50 indicate that CFI bias was similar across estimation methods with the weighted and generalized least squares estimators producing more negatively biased results when the number of clusters equaled 30 and cluster size was held at 10, while the maximum likelihood estimator produced negligible bias. As sample size increased from 300, each of the estimation methods produced similar results with little to no bias reported. Mean CFI bias ranged from -1.033 (*SD* = 13.183) for the weighted least squares estimator when the number of clusters was 30 and cluster size was 10 in severely nonnormal data to .009 (*SD* = .274) for the maximum likelihood estimator when the number of clusters was 30 and cluster size was 10 in moderately nonnormal dataconditions. Note the least amount of bias related to the CFI was associated with the maximum likelihood estimator across all normal and nonnormal data conditions investigated in the present study.

Table 50
Descriptive Results of CFI Bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	-.008	.194	-.306	1.405	-.328	1.546
	N = 900 (30 x 30)	.000	.024	-.018	.079	-.013	.048
	N = 1500 (30 x 50)	.001	.006	-.001	.006	.000	.004
	N = 500 (50 x 10)	.003	.006	-.004	.020	.001	.010
	N = 1500 (50 x 30)	.002	.003	-.002	.010	.000	.007
	N = 2500 (50 x 50)	.002	.002	.001	.004	.002	.002
	N = 1000 (100 x 10)	.000	.006	-.002	.008	.000	.005
	N = 3000 (100 x 30)	.001	.001	.001	.002	.001	.002
	N = 5000 (100 x 50)	.001	.002	.001	.002	.001	.001
Moderately Nonnormal	N = 300 (30 x 10)	.009	.274	-.113	.909	-.079	.739
	N = 900 (30 x 30)	-.003	.036	-.031	.252	-.021	.134
	N = 1500 (30 x 50)	.001	.008	-.005	.056	-.003	.026
	N = 500 (50 x 10)	-.003	.080	-.021	.123	-.013	.088
	N = 1500 (50 x 30)	.002	.004	-.002	.016	.000	.008
	N = 2500 (50 x 50)	.002	.002	.001	.004	.002	.003
	N = 1000 (100 x 10)	-.001	.015	-.006	.044	-.002	.021
	N = 3000 (100 x 30)	.001	.014	-.001	.033	.001	.014
	N = 5000 (100 x 50)	.001	.002	.001	.002	.001	.001
Severely Nonnormal	N = 300 (30 x 10)	-.003	.635	-.687	7.466	-1.033	13.183
	N = 900 (30 x 30)	-.003	.099	-.081	.754	-.049	.459
	N = 1500 (30 x 50)	.000	.012	-.004	.059	-.001	.023
	N = 500 (50 x 10)	-.016	.229	-.182	1.073	-.079	.501
	N = 1500 (50 x 30)	.001	.004	.000	.006	.000	.006
	N = 2500 (50 x 50)	.000	.022	-.002	.035	.000	.009
	N = 1000 (100 x 10)	.000	.007	-.006	.025	-.005	.025
	N = 3000 (100 x 30)	.001	.002	.000	.005	.001	.003
	N = 5000 (100 x 50)	.001	.003	.000	.008	.001	.003

The results reported in Table 51 indicated that bias among the root mean square residual (RMR) was similar across estimation methods in normal data conditions. As data increased in departure from normality, the weighted least squares estimator produced slightly less biased RMR results in severely nonnormal data when compared to the maximum likelihood and generalized least squares estimators. Note the standard deviation associated with RMR bias among the weighted least squares estimates was greater in magnitude indicating that there was increased instability in the RMR bias associated with the weighted least squares estimator. The greatest amount of RMR bias was noted when cluster size was greater than the number of clusters (i.e., when the number of clusters equaled 30 and cluster size was held at 50). Average RMR bias ranged from -.391 (SD = .466) with the maximum likelihood estimator in normal data conditions when the number of clusters equaled 30 and cluster size was 50 to m = .301 (SD = 1.231) for the weighted least squares estimator in normal data conditions when the sample size equaled 300 (30 clusters with cluster size of 10).

Table 51
Descriptive Results of RMR Bias among Selected Design Conditions

Normality	Sample Size (No. Clusters x Cluster Size)	Estimation Method					
		ML		GLS		WLS	
		M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	.155	1.047	.273	1.107	.301	1.231
	N = 900 (30 x 30)	-.096	.700	-.038	.769	-.096	.690
	N = 1500 (30 x 50)	-.391	.466	-.360	.467	-.381	.472
	N = 500 (50 x 10)	-.051	.727	-.040	.738	.003	.793
	N = 1500 (50 x 30)	.042	.826	.079	.891	.073	.839
	N = 2500 (50 x 50)	-.248	.657	-.212	.701	-.253	.622
	N = 1000 (100 x 10)	-.239	.616	-.263	.573	-.264	.601
	N = 3000 (100 x 30)	.086	.846	.048	.789	.090	.873
	N = 5000 (100 x 50)	.060	.788	.107	.821	.128	.850
Moderately Nonnormal	N = 300 (30 x 10)	.105	1.123	.294	2.225	.270	1.895
	N = 900 (30 x 30)	-.183	.665	-.155	.708	-.171	.686
	N = 1500 (30 x 50)	-.367	.502	-.352	.515	-.366	.509
	N = 500 (50 x 10)	.139	.954	.143	.963	.116	.933
	N = 1500 (50 x 30)	-.014	.792	-.034	.764	-.031	.760
	N = 2500 (50 x 50)	-.016	.750	-.022	.733	-.019	.764
	N = 1000 (100 x 10)	-.305	.577	-.317	.553	-.306	.535
	N = 3000 (100 x 30)	.002	.765	-.021	.745	-.041	.744
	N = 5000 (100 x 50)	.039	.787	.055	.807	.045	.798
Severely Nonnormal	N = 300 (30 x 10)	-.031	.777	.082	1.372	.218	2.070
	N = 900 (30 x 30)	-.219	.609	-.199	.624	-.208	.633
	N = 1500 (30 x 50)	-.389	.481	-.374	.487	-.383	.489
	N = 500 (50 x 10)	-.050	.734	-.035	.756	.012	.872
	N = 1500 (50 x 30)	-.085	.702	-.072	.696	-.100	.687
	N = 2500 (50 x 50)	-.035	.766	-.025	.783	-.025	.772
	N = 1000 (100 x 10)	-.323	.514	-.329	.516	-.308	.539
	N = 3000 (100 x 30)	.044	.820	-.004	.774	.019	.799
	N = 5000 (100 x 50)	.049	.802	.056	.801	.036	.815

Bias among AIC reported in Table 52 was varied in magnitude in small sample sizes, especially when the number of clusters was 30 and cluster size ranged from 10 to 30. In these same conditions (sample size = 300 and 900), the weighted least squares estimator generally produced less biased results in each of the data normality conditions examined when compared to the maximum likelihood and generalized least squares estimators. Mean AIC bias ranged from -.737 (*SD* = .678) among the maximum likelihood estimator in normal data conditions in 50 clusters with a cluster size of 10 to 57.848 (*SD* =644.86) among the weighted least squares estimator in severely nonnormal data in 30 clusters with cluster size held at 10.

In small sample sizes, AIC bias was positive for each estimation method. As sample size increased, the weighted least squares estimator produced negatively biased results while positive bias was associated with the maximum likelihood and generalized least squares estimators.

Table 52
Descriptive Results of AIC Bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	23.261	141.487	14.216	62.380	22.988	100.821
	N = 900 (30 x 30)	1.765	6.728	1.888	5.407	1.227	5.131
	N = 1500 (30 x 50)	-.601	.725	-.579	.680	-.763	.417
	N = 500 (50 x 10)	-.737	.678	-.742	.605	-.821	.488
	N = 1500 (50 x 30)	.077	2.436	.283	3.084	-.095	2.458
	N = 2500 (50 x 50)	-.414	1.185	-.362	1.361	-.631	.806
	N = 1000 (100 x 10)	-.581	.807	-.619	.729	-.651	.848
	N = 3000 (100 x 30)	.041	1.746	-.056	1.554	-.186	1.830
	N = 5000 (100 x 50)	.180	1.871	.255	2.082	-.220	1.319
Moderately Nonnormal	N = 300 (30 x 10)	12.561	242.992	7.533	58.786	8.539	70.434
	N = 900 (30 x 30)	14.538	138.104	10.005	63.015	6.797	40.524
	N = 1500 (30 x 50)	-.107	2.852	-.071	3.295	-.256	2.699
	N = 500 (50 x 10)	.725	16.470	-.012	4.536	.055	5.596
	N = 1500 (50 x 30)	.740	6.757	.605	5.828	-.003	3.208
	N = 2500 (50 x 50)	-.394	1.595	-.430	1.340	-.619	.955
	N = 1000 (100 x 10)	-.238	7.176	-.397	3.275	-.467	2.689
	N = 3000 (100 x 30)	.020	2.036	-.039	2.040	-.361	1.651
	N = 5000 (100 x 50)	.100	2.095	.216	2.600	-.265	1.766
Severely Nonnormal	N = 300 (30 x 10)	25.200	699.772	31.909	343.486	57.848	644.186
	N = 900 (30 x 30)	9.152	178.516	6.242	52.432	3.937	37.846
	N = 1500 (30 x 50)	-.114	3.472	-.085	3.440	-.337	4.358
	N = 500 (50 x 10)	10.301	166.135	5.142	34.187	3.035	24.538
	N = 1500 (50 x 30)	-.223	1.449	-.212	1.441	-.475	1.277
	N = 2500 (50 x 50)	-.397	1.440	-.376	1.532	-.644	.963
	N = 1000 (100 x 10)	-.225	2.973	-.286	2.884	-.190	3.002
	N = 3000 (100 x 30)	.141	3.646	.090	4.302	-.365	1.827
	N = 5000 (100 x 50)	.185	2.125	.221	2.264	-.286	1.763

Bivariate Results

Pearson product-moment correlations were calculated to determine how bias among the selected fit indices was impacted by the number of clusters, cluster size, correlation among variables, and data normality conditions. The results displayed in Table 53 show that among the study design conditions, statistically significant correlations ranged from r = -.029 between cluster size and the fit index (95% CI ranged from -.033 to -.024) to r = .025 between cluster size and the goodness of fit index (95% CI ranged from .020 to .029). Although the majority of the correlations among the study design conditions and fit indices were statistically significant, squaring the correlation coefficients resulted in effect sizes that explained less than one percent of the variation in the outcome variable. A plausible explanation for the negligible correlations could be due to the simplistic model (model one) that included only one predictor at each level and a cross-level interaction term.

Table 53

Pearson Product-Moment Correlations among Parameter Fit Indices Bias and Selected Design Conditions

	1	2	3	4	5	6	7	8	9	10
GFI Bias (1)	1.00									
CFI Bias (2)	.960**	1.00								
Fit Bias (3)	-.670**	-.706**	1.00							
RMR Bias (4)	-.269**	-.263**	.224**	1.00						
AIC Bias (5)	-.675**	-.712**	.970**	.225**	1.00					
No. of Clusters (6)	.021**	.020**	-.023**	.010**	-.026**	1.00				
Cluster Size (7)	.025**	.024**	-.029**	-.041**	-.026**	-.004	1.00			
Correlation among Variables (8)	-.003	-.002	.010**	.018**	.006*	.000	.000	1.00		
Skew (9)	-.014**	-.015**	.007**	-.002	.007**	.001	.007**	.000	1.00	
Kurtosis (10)	-.021**	-.022**	.018**	-.025**	.016**	.000	.006*	.000	.396**	1.00

**.Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed). The sample size for each coefficient was 181,875.

Factorial Analysis Results of Fit Indices

Table 54 displays the results of GFI bias by the six study design conditions for model one. Each of the design conditions as main and interaction effects was statistically significant. However, as indicated by the partial eta square results, each of the statistically significant effects explained less than one percent of the variance in the outcome variable. To provide further insight into the results, post hoc confidence intervals were calculated that included the number of clusters, cluster size, and estimation method. The results are displayed in Figure 15.

Based on the results in Figure 15, the number of clusters and cluster size appeared to be the most important factors impacting GIF bias. In small sample sizes, especially when the number of clusters equaled 30 and cluster size was held at 10 and 30, each estimation method produced negatively biased results, with the greatest amount of negative bias associated with the weighted and generalized least squares estimators. Similar results were found when the number of clusters equaled 50 and cluster size was held at 10. As sample size increased, GFI bias decreased with each estimation method producing little to no bias.

Table 54
Factorial ANOVA Results of the Study's Six Design Conditions Effect on GFI Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	848.365	88	9.641	23.169	< .001	.011[a]
Intercept	27.095	1	27.095	65.119	< .001	.000
Est. Mthd	10.856	2	5.428	13.045	< .001	.000
Clstrs	28.682	2	14.341	34.466	< .001	.000
Clstr. Size	42.515	2	21.257	51.088	< .001	.001
Corr	17.692	2	8.846	21.260	< .001	.000
Skew	10.499	3	3.500	8.411	< .001	.000
Kurt	33.895	3	11.298	27.154	< .001	.000
Est. Mthd. * Clstrs	25.854	4	6.463	15.534	< .001	.000
Est. Mthd. * Clstr. Size	28.035	4	7.009	16.844	< .001	.000
Est. Mthd. * Corr	17.915	4	4.479	10.764	< .001	.000
Est. Mthd. * Skew	9.235	6	1.539	3.699	< .001	.000
Est. Mthd. * Kurt	25.340	6	4.223	10.150	< .001	.000
Clstrs * Clstr. Size	79.092	4	19.773	47.521	< .001	.001
Clstrs * Corr	43.936	4	10.984	26.398	< .001	.001
Clstrs * Skew	19.875	6	3.312	7.961	< .001	.000
Clstrs * Kurt	67.774	6	11.296	27.147	< .001	.001
Clstr. Size * Corr	46.047	4	11.512	27.667	< .001	.001
Clstr. Size * Skew	29.000	6	4.833	11.616	< .001	.000
Clstr. Size * Kurt	50.449	6	8.408	20.208	< .001	.001
Corr * Skew	38.501	6	6.417	15.422	< .001	.001
Corr * Kurt	81.035	6	13.506	32.459	< .001	.001
Skew * Kurt	15.546	2	7.773	18.682	< .001	.000
Error	75221.923	180783	.416			

Note. The model eta squared for the full factorial 6-way model was .051 with df=696. The model eta squared for only the main and the two-way interaction effects was .011 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .040 (i.e., .051-.011).

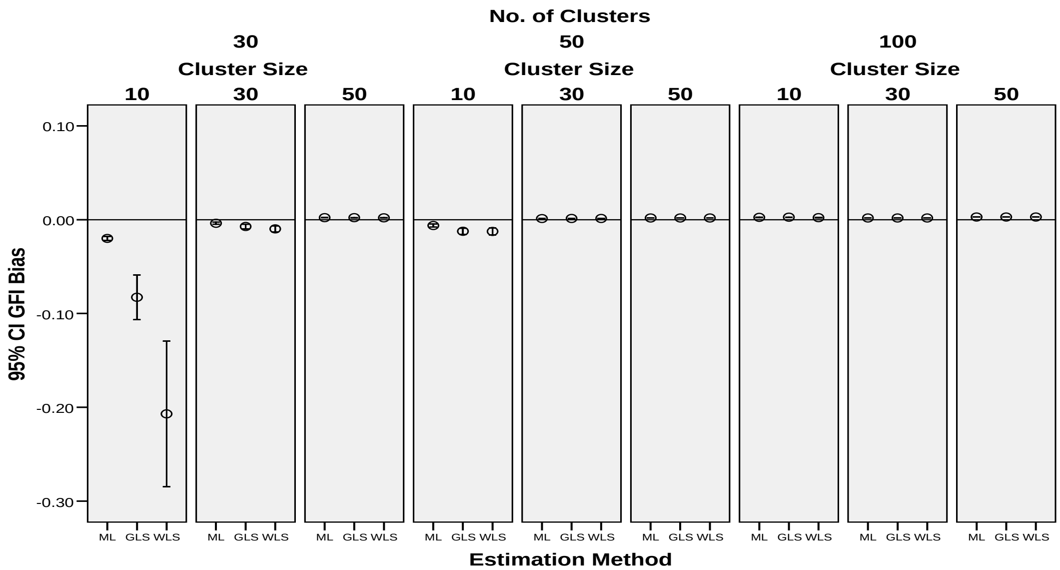

Figure 15. Confidence intervals examining bias among the GFI by estimation method, cluster size, and the number of clusters.

Table 55 displays the results of the study’s design conditions on fit bias. The results revealed that the number of clusters, cluster size, correlation amongvariables, and kurtosis had a statistically significant impact on fit bias as main effects, while the estimation method and skew were not statistically significant. Regarding interaction effects, estimation method by the number of clusters and estimation method by cluster size were not statistically significant effects while the remaining interaction terms shared a statistically significant relation with the outcome variable. Similar to GFI bias, the resulting partial eta square indicated that each of the statistically significant effects explained less than one percent of the variance in fit bias.

Post hoc confidence intervals were derived to examine fit bias by the number of clusters, cluster size, correlation among variables and kurtosis. The results are presented in Figure 16.

Table 55

Factorial ANOVA Results of the Study's Six Design Conditions Effect on Fit Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	33783407	88	383902	17.91	< .001	.009[a]
Intercept	2123250	1	2E+006	99.06	< .001	.001
Est. Mthd	24920.39	2	12460	.581	.559	.000
Clstrs	1419313	2	709656	33.11	< .001	.000
Clstr. Size	2962896	2	1E+006	69.12	< .001	.001
Corr	733948.3	2	366974	17.12	< .001	.000
Skew	143536.1	3	47845	2.232	.082	.000
Kurt	1371963	3	457321	21.34	< .001	.000
Est. Mthd. * Clstrs	199402.9	4	49851	2.326	.054	.000
Est. Mthd. * Clstr. Size	150808.2	4	37702	1.759	.134	.000
Est. Mthd. * Corr	337124.2	4	84281	3.932	.003	.000
Est. Mthd. * Skew	535297.8	6	89216	4.163	< .001	.000
Est. Mthd. * Kurt	516115.0	6	86019	4.013	.001	.000
Clstrs * Clstr. Size	4022631	4	1E+006	46.92	< .001	.001
Clstrs * Corr	2201489	4	550372	25.68	< .001	.001
Clstrs * Skew	764648.9	6	127441	5.946	< .001	.000
Clstrs * Kurt	3962651	6	660442	30.81	< .001	.001
Clstr. Size * Corr	1588897	4	397224	18.53	< .001	.000
Clstr. Size * Skew	345183.4	6	57531	2.684	.013	.000
Clstr. Size * Kurt	1780312	6	296719	13.84	< .001	.000
Corr * Skew	1330742	6	221790	10.35	< .001	.000
Corr * Kurt	4073115	6	678852	31.67	< .001	.001
Skew * Kurt	471797.2	2	235899	11.01	< .001	.000
Error	3.9E+009	180783	21433			
Total	3.9E+009	180872				
Corrected Total	3.9E+009	180871				

a. R Squared = .009 (Adjusted R Squared = .008)

Note. The model eta squared for the full factorial 6-way model was .033 with df=696. The model eta squared for only the main and the two-way interaction effects was .009 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .024 (i.e., .033-.009).

The results displayed in Figure 16 indicated that FIT bias was somewhat erratic in small sample sizes, especially when the number of clusters was held at 30 and 50, cluster size was 10, and kurtosis ranged from one to seven. The results stabilized across each of the conditions investigated when the sample size increased from 900 to 5,000 with little to no bias noted among the conditions examined.

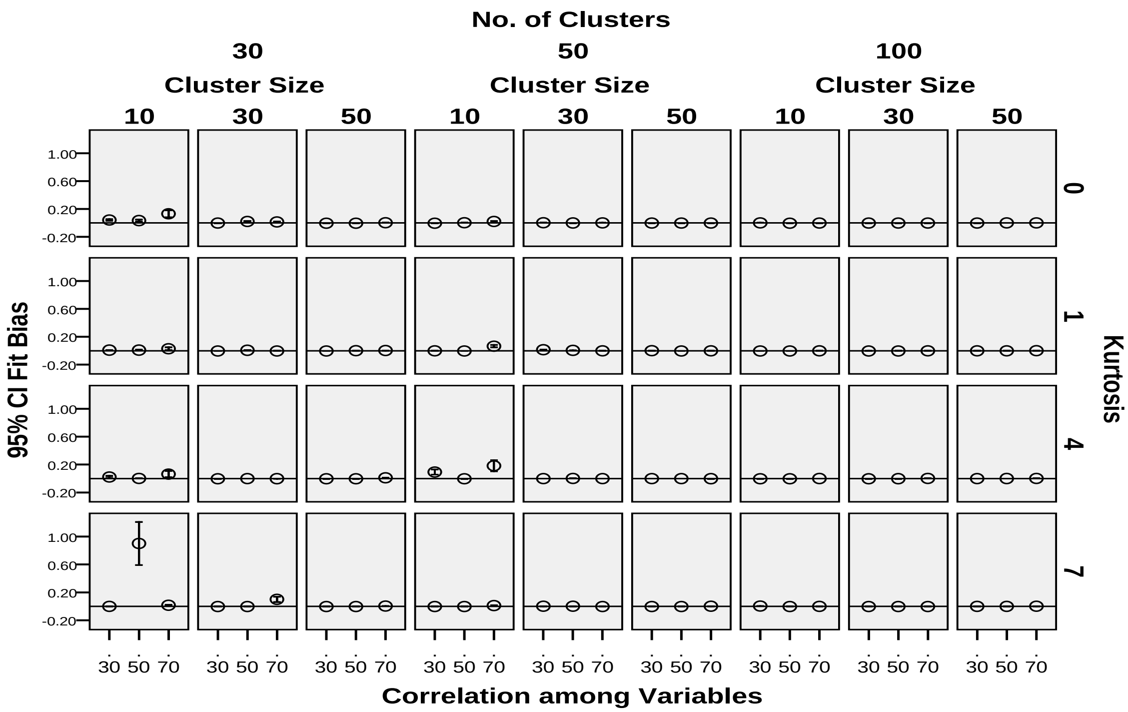

Figure 16. Confidence intervals examining bias among the Fit index by the correlation among variables, the number of clusters, cluster size, and kurtosis.

The results reported in Table 56 regarding comparative fit index bias indicated that each of the design conditions as main effects were statistically significant. However, similar to the results reported for the prior fit indices, each statistically significant main effect explained less than one percent of the variation in the outcome variable. Similar results were found for the interaction terms with the exception of estimation method by skew, which was not statistically significant. Consistent with the prior fit indices for model one, each statistically significant interaction effect explained less than one percent of the variation in bias associated with the comparative fit index.

Figure 17 displays the post hoc confidence intervals examining CFI bias by the number of clusters, cluster size, kurtosis, and estimation method. The results revealed that severe kurtosis (kurtosis = 7) had a negative impact on CFI bias in small sample sizes, especially when the number of clusters was 30 and cluster membership was 10 for both the generalized and weighted least squares estimators. In this scenario, CFI bias was negligible for the maximum likelihood estimator. Although less in magnitude, similar results were found when the sample size increased to 500 and kurtosis was held at 4. As sample size increased beyond 500, the CFI bias was negligible with similar results reported across estimation methods in varying degrees of kurtosis.

Table 56
Factorial ANOVA Results of the Study's Six Design Conditions Effect on CFI Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	8258.407	88	93.846	24.26	< .001	.012[a]
Intercept	332.823	1	332.82	86.04	< .001	.000
Est. Mthd	103.049	2	51.524	13.32	< .001	.000
Clstrs	242.753	2	121.38	31.38	< .001	.000
Clstr. Size	392.082	2	196.04	50.68	< .001	.001
Corr	195.656	2	97.828	25.29	< .001	.000
Skew	101.547	3	33.849	8.750	< .001	.000
Kurt	362.747	3	120.92	31.26	< .001	.001
Est. Mthd. * Clstrs	224.917	4	56.229	14.54	< .001	.000
Est. Mthd. * Clstr. Size	254.421	4	63.605	16.44	< .001	.000
Est. Mthd. * Corr	114.106	4	28.526	7.374	< .001	.000
Est. Mthd. * Skew	56.824	6	9.471	2.448	.023	.000
Est. Mthd. * Kurt	170.037	6	28.339	7.326	< .001	.000
Clstrs * Clstr. Size	638.664	4	159.67	41.28	< .001	.001
Clstrs * Corr	451.264	4	112.82	29.16	< .001	.001
Clstrs * Skew	194.681	6	32.447	8.388	< .001	.000
Clstrs * Kurt	724.199	6	120.70	31.20	< .001	.001
Clstr. Size * Corr	550.693	4	137.67	35.59	< .001	.001
Clstr. Size * Skew	282.113	6	47.019	12.15	< .001	.000
Clstr. Size * Kurt	551.567	6	91.928	23.76	< .001	.001
Corr * Skew	387.153	6	64.526	16.68	< .001	.001
Corr * Kurt	862.405	6	143.73	37.16	< .001	.001
Skew * Kurt	155.752	2	77.876	20.13	< .001	.000
Error	699079.9	180719	3.868			
Total	707660.8	180808				
Corrected Total	707338.3	180807				

a. R Squared = .012 (Adjusted R Squared = .011)

Note. The model eta squared for the full factorial 6-way model was .049 with df=696. The model eta squared for only the main and the two-way interaction effects was .012 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .037 (i.e., .049-.012).

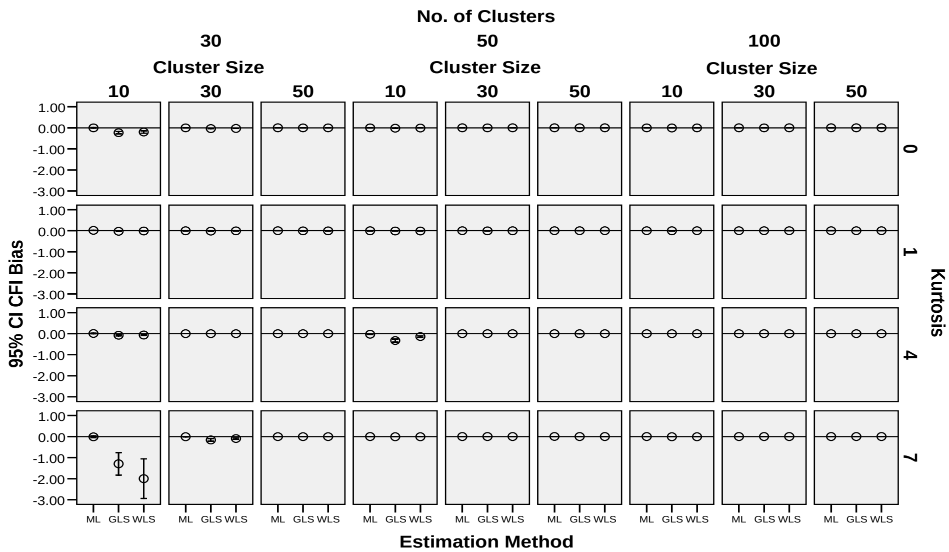

Figure 17. Confidence intervals examining CFI bias by the estimation method, kurtosis, the number of clusters and cluster size

The results displayed in Table 57 reveal that each of the design conditions as main and interaction effects were statistically significant factors impacting RMR bias. Regarding main effects, the number of clusters explained the greatest amount of variation in the outcome variable (partial $\eta^2 = .004$). As for interaction effects, the number of clusters by cluster size explained more than three percent of the variation in the dependent variable (partial $\eta^2 = .032$), while the remaining terms explained less than one percent of the variance in the criterion variable. Post hoc confidence intervals displayed in Figure 18 show the comparison of the salient statistically significant effects.

Table 57
Factorial ANOVA Results of the Study's Six Design Conditions Effect on RMR Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	7686.604	88	87.348	119.172	< .001	.055[a]
Intercept	999.808	1	999.808	1364.074	< .001	.007
Est. Mthd	6.775	2	3.388	4.622	.010	.000
Clstrs	525.374	2	262.687	358.393	< .001	.004
Clstr. Size	61.313	2	30.657	41.826	< .001	.000
Corr	8.446	2	4.223	5.761	.003	.000
Skew	187.843	3	62.614	85.427	< .001	.001
Kurt	270.550	3	90.183	123.041	< .001	.002
Est. Mthd. * Clstrs	42.129	4	10.532	14.370	< .001	.000
Est. Mthd. * Clstr. Size	41.567	4	10.392	14.178	< .001	.000
Est. Mthd. * Corr	8.218	4	2.054	2.803	.024	.000
Est. Mthd. * Skew	29.902	6	4.984	6.799	< .001	.000
Est. Mthd. * Kurt	20.571	6	3.429	4.678	< .001	.000
Clstrs * Clstr. Size	4427.910	4	1106.978	1510.290	< .001	.032
Clstrs * Corr	397.576	4	99.394	135.607	< .001	.003
Clstrs * Skew	208.586	6	34.764	47.430	< .001	.002
Clstrs * Kurt	251.731	6	41.955	57.241	< .001	.002
Clstr. Size * Corr	201.060	4	50.265	68.579	< .001	.002
Clstr. Size * Skew	344.325	6	57.387	78.296	< .001	.003
Clstr. Size * Kurt	228.813	6	38.135	52.030	< .001	.002
Corr * Skew	246.179	6	41.030	55.979	< .001	.002
Corr * Kurt	139.280	6	23.213	31.671	< .001	.001
Skew * Kurt	128.410	2	64.205	87.597	< .001	.001
Error	132506.140	180783	.733			
Total	141238.772	180872				
Corrected Total	140192.744	180871				

a. R Squared = .055 (Adjusted R Squared = .054)

Note. The model eta squared for the full factorial 6-way model was .114 with df=696. The model eta squared for only the main and the two-way interaction effects was .055 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .055 (i.e., .114-.055).

The results displayed in Figure 18 show that as kurtosis increased, RMR bias variation increased in small and moderately small sample sizes. As sample size increased, namely, as the number of clusters increased to 50 and 100 and cluster size was at least 30, RMR bias decreased.

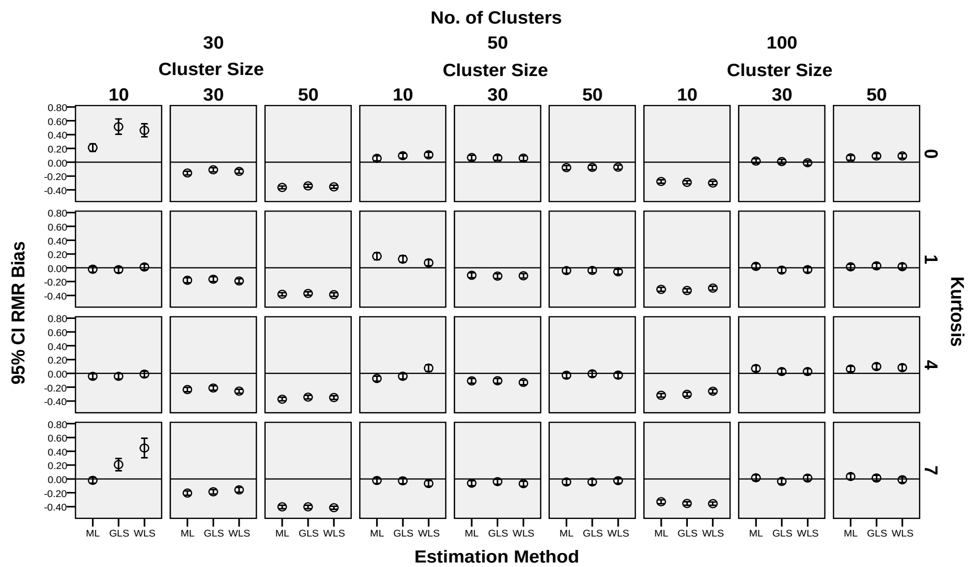

Figure 18. Confidence intervals examining RMR bias by estimation method, the number of clusters, cluster size, and kurtosis

The results of AIC bias reported in Table 58 revealed that each of the study design conditions, with the exception of estimation method, had a statistically significant impact on AIC bias. However, each statistically significant main effect explained less than one percent of the variation in the outcome variable. Regarding interaction terms, the term comprised of estimation method by the number of clusters was not statistically significant, while the remaining interaction effects were statistically significant. Similar to the main effects, each of the statistically significant interaction terms explained less than one percent of the variation in AIC bias. The largest partial eta square was associated with the number of clusters as a main effect and the interaction term that included the number of clusters by cluster size as an interaction term (partial $\eta^2 = .001$).

Table 58
Factorial ANOVA Results of the Study's Six Design Conditions Effect on AIC Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	3E+007	88	360371	18.68	< .001	.009[a]
Intercept	1927315	1	2E+006	99.92	< .001	.001
Est. Mthd	17260.4	2	8630.2	.447	.639	.000
Clstrs	1893339	2	946670	49.08	< .001	.001
Clstr. Size	1642396	2	821198	42.57	< .001	.000
Corr	758230	2	379115	19.65	< .001	.000
Skew	374277	3	124759	6.468	< .001	.000
Kurt	1702453	3	567484	29.42	< .001	.000
Est. Mthd. * Clstrs	103893	4	25973	1.347	.250	.000
Est. Mthd. * Clstr. Size	214159	4	53540	2.776	.025	.000
Est. Mthd. * Corr	212601	4	53150	2.755	.026	.000
Est. Mthd. * Skew	450050	6	75008	3.889	.001	.000
Est. Mthd. * Kurt	363934	6	60656	3.145	.004	.000
Clstrs * Clstr. Size	3514344	4	878586	45.55	< .001	.001
Clstrs * Corr	2799975	4	699994	36.29	< .001	.001
Clstrs * Skew	721406	6	120234	6.233	< .001	.000
Clstrs * Kurt	3695991	6	615998	31.94	< .001	.001
Clstr. Size * Corr	1418274	4	354568	18.38	< .001	.000
Clstr. Size * Skew	770988	6	128498	6.662	< .001	.000
Clstr. Size * Kurt	1830009	6	305002	15.81	< .001	.001
Corr * Skew	1054083	6	175680	9.108	< .001	.000
Corr * Kurt	3260853	6	543475	28.18	< .001	.001
Skew * Kurt	339988	2	169994	8.813	< .001	.000
Error	3E+009	180783	19289			
Total	4E+009	180872				
Corrected Total	4E+009	180871				

a. R Squared = .009 (Adjusted R Squared = .009)

Note. The model eta squared for the full factorial 6-way model was .034 with df=696. The model eta squared for only the main and the two-way interaction effects was .009 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .025 (i.e., .034-.009).

Figure 19 displays the post hoc confidence intervals examining AIC bias by the number of clusters, cluster size, correlation among variables, and kurtosis. In normal data conditions (kurtosis = 0), AIC bias varied considerably when the number of clusters equaled 30 and cluster size was held at 10 and the correlation among variables was at .70. In addition, as kurtosis increased to 7 in the same sample size, AIC bias increased sharply when the correlation among variables was .50. Slight positive bias was noted among varying degrees of kurtosis when the number of clusters was 50 and cluster size was 10. AIC bias was negligible when the number of clusters was 50 and cluster size was 30 with similar results reported for 100 clusters with cluster size ranging from 10 to 50. In other words, as sample size increased to 1500 and above, AIC bias was negligible.

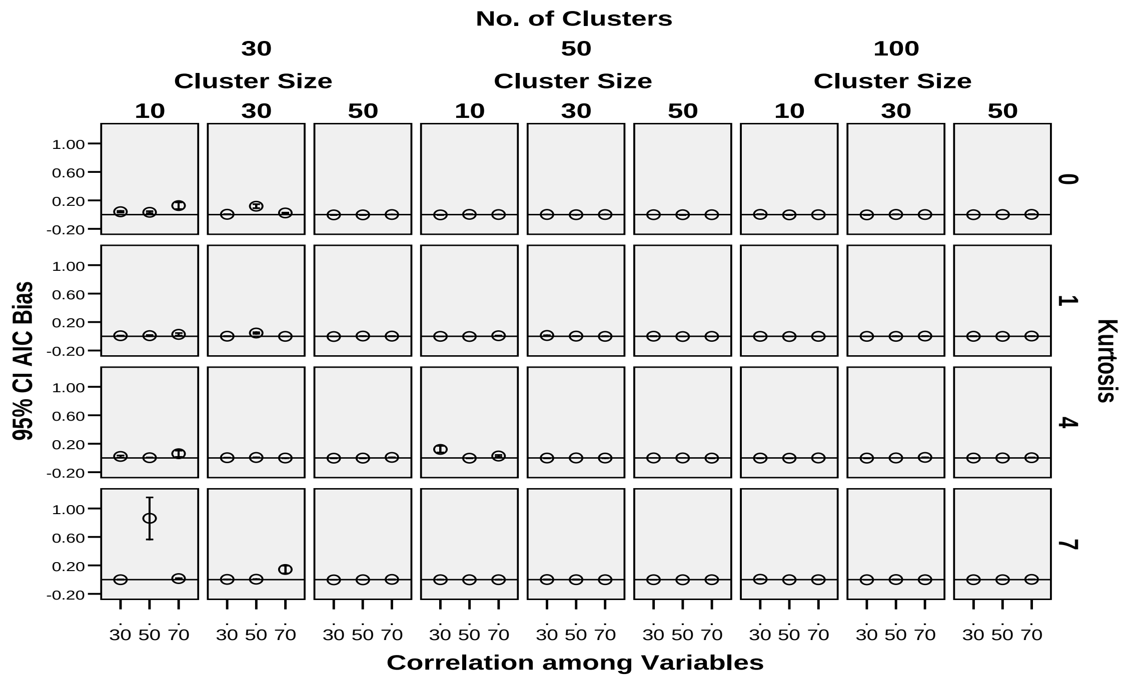

Figure 19. Confidence intervals examining AIC bias by the correlation among variables, the number of clusters, cluster size, and kurtosis.

Summary Model One

The results from model one indicated that of the six conditions examined, sample size appeared to have the greatest impact on the goodness of fit index, while small sample size (less than 900) and severe kurtosis (kurtosis = 7) impacted the fit, AIC and the comparative fit index. This was especially evident when the sample size was less than 900.

The results of the size study conditions regarding bias among the root mean square residual revealed that the number of clusters by cluster size had the greatest impact on RMR bias (partial η^2 =.032). As kurtosis increased from one to seven, the variation in RMR bias increased in small and moderately small sample sizes. As sample size increased, namely, as the number of clusters increased to 50 and 100 and cluster size was at least 30, RMR bias decreased sharply across estimation methods when the sample size was 3,000 and 5,000 respectively.

Model Two

Model two included two level-1 and one level-2 predictor variables and a cross-level interaction term between the initial level-1 and the level-2 variable. The model is displayed below.

Level-1 equation $Y_{ij} = \beta_0 + \beta_1 X1_{ij} + \beta_2 \; X2_{ij} + \varepsilon_{ij}$

Level-2 equation $\beta_0 = \gamma 00 + U0_j$

$\beta_1 = \gamma 10 + \gamma_{11} \; W_1 BAR$

$\beta_2 = \gamma 20 + \gamma 21 \; W_2 BAR$

Combined equation $Y = \gamma_{00} + \gamma_0 X1 + \gamma_{11} \; W_1 BAR{*}X1 + \gamma_{20} \; x_4 +$

$\gamma_{21}{*} \; W_2 BAR{*}X2 \; U_0 + \varepsilon$

Descriptive Measures of Bias among Selected Fit Indices

The descriptive results of GFI bias presented in Table 59 for model two showed that bias was negligible across estimation methods in each of the data normality and sample sizes investigated. Comparing the results to those reported for model one, it is apparent that in small sample sizes for the simple model one, there was variation in GFI bias, especially when the number of clusters was 30 and cluster size was held at 10 in each data normality condition and across estimation methods. However, in the more complex model two, the amount of variation in GFI bias was negligible in each of the design conditions displayed in Table 56. Mean GFI bias ranged from .000 (SD = .002) for the maximum likelihood estimator in normal and moderately nonnormal data conditions when the number of clusters equaled 30 and cluster size was 10 to .003 (*SD* = .000) for the weighted least squares estimator in normal and moderately nonnormal data conditions when the number of clusters and cluster size was 50.

Table 59
Descriptive Results of GFI Bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	.000	.002	.000	.003	.001	.001
	N = 900 (30 x 30)	.001	.001	.001	.001	.001	.001
	N = 1500 (30 x 50)	.001	.001	.001	.001	.001	.001
	N = 500 (50 x 10)	.002	.002	.002	.002	.002	.001
	N = 1500 (50 x 30)	.002	.001	.002	.001	.002	.000
	N = 2500 (50 x 50)	.002	.001	.002	.001	.003	.000
	N = 1000 (100 x 10)	.001	.001	.001	.001	.001	.000
	N = 3000 (100 x 30)	.001	.000	.001	.000	.001	.000
	N = 5000 (100 x 50)	.002	.000	.002	.000	.002	.000
Moderately Nonnormal	N = 300 (30 x 10)	.000	.002	.000	.003	.001	.001
	N = 900 (30 x 30)	.001	.001	.001	.001	.001	.001
	N = 1500 (30 x 50)	.001	.001	.001	.001	.001	.001
	N = 500 (50 x 10)	.002	.001	.002	.002	.002	.001
	N = 1500 (50 x 30)	.002	.000	.002	.000	.002	.000
	N = 2500 (50 x 50)	.002	.001	.002	.001	.003	.000
	N = 1000 (100 x 10)	.000	.001	.001	.001	.001	.000
	N = 3000 (100 x 30)	.001	.000	.001	.000	.001	.000
	N = 5000 (100 x 50)	.002	.000	.002	.000	.002	.000
Severely Nonnormal	N = 300 (30 x 10)	.000	.003	.000	.003	.000	.003
	N = 900 (30 x 30)	.001	.001	.001	.001	.001	.001
	N = 1500 (30 x 50)	.001	.001	.001	.001	.001	.001
	N = 500 (50 x 10)	.002	.002	.002	.002	.002	.002
	N = 1500 (50 x 30)	.002	.000	.002	.001	.002	.001
	N = 2500 (50 x 50)	.002	.001	.002	.001	.002	.001
	N = 1000 (100 x 10)	.001	.001	.000	.001	.001	.001
	N = 3000 (100 x 30)	.001	.000	.001	.000	.001	.000
	N = 5000 (100 x 50)	.002	.000	.002	.000	.002	.000

Table 60 displays Fit bias among varying data normality conditions, sample size, and estimation methods. The results indicate that the greatest amount of bias associated with the fit index was noted across estimation methods when cluster size was greater than the number of clusters (i.e., clusters = 30, cluster size = 50). Mean fit bias ranged from -.733 (*SD* = .379) for the weighted least squares estimator in moderately nonnormal data conditions when the number of clusters was 30 and cluster size was 50 to .252 (*SD* = 1.946) with the maximum likelihood estimator in severely nonnormal data conditions when the number of clusters equaled 30 and cluster membership was held constant at 10. In each of the data conditions investigated in the present study, the weighted least squares produced negatively biased results in each of the data normality conditions across all sample sizes. Compared to the maximum likelihood estimator, the weighted least squares estimator produced negatively biased results that were greater in magnitude (almost twice as large). Similar results were reported for the both the maximum likelihood and generalized least squares estimators.

Table 60
Descriptive Results of Fit Bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	.104	1.522	.166	1.769	-.466	.795
	N = 900 (30 x 30)	.059	1.570	.240	1.761	-.585	.621
	N = 1500 (30 x 50)	-.319	.964	-.330	.964	-.747	.345
	N = 500 (50 x 10)	.094	1.475	-.015	1.467	-.524	.668
	N = 1500 (50 x 30)	-.281	1.003	-.331	1.001	-.756	.327
	N = 2500 (50 x 50)	-.083	1.313	-.025	1.387	-.648	.494
	N = 1000 (100 x 10)	-.044	1.282	-.032	1.423	-.556	.676
	N = 3000 (100 x 30)	-.228	1.406	-.170	1.336	-.728	.463
	N = 5000 (100 x 50)	-.044	1.417	.017	1.418	-.618	.524
Moderately Nonnormal	N = 300 (30 x 10)	.133	1.603	.153	1.693	-.447	.896
	N = 900 (30 x 30)	.159	1.669	.156	1.629	-.549	.650
	N = 1500 (30 x 50)	-.328	.944	-.323	.953	-.733	.379
	N = 500 (50 x 10)	-.023	1.401	.080	1.578	-.486	.733
	N = 1500 (50 x 30)	-.356	.917	-.312	.974	-.731	.397
	N = 2500 (50 x 50)	.034	1.464	-.026	1.338	-.618	.539
	N = 1000 (100 x 10)	.030	1.437	-.002	1.427	-.549	.660
	N = 3000 (100 x 30)	-.234	1.332	-.215	1.378	-.680	.567
	N = 5000 (100 x 50)	.022	1.534	.006	1.364	-.614	.553
Severely Nonnormal	N = 300 (30 x 10)	.252	1.946	.189	1.732	-.233	1.209
	N = 900 (30 x 30)	.161	1.670	.209	1.712	-.419	.898
	N = 1500 (30 x 50)	-.306	.997	-.292	1.003	-.707	.447
	N = 500 (50 x 10)	.018	1.455	.041	1.607	-.315	1.056
	N = 1500 (50 x 30)	-.308	1.000	-.284	1.008	-.666	.513
	N = 2500 (50 x 50)	.031	1.509	.041	1.477	-.566	.687
	N = 1000 (100 x 10)	-.021	1.374	.041	1.479	-.370	.990
	N = 3000 (100 x 30)	-.207	1.461	-.244	1.362	-.621	.725
	N = 5000 (100 x 50)	.052	1.494	.059	1.483	-.550	.748

The descriptive results of CFI bias by data normality condition, sample size, and estimation methods are shown in Table 61. CFI bias was negligible in each of the conditions examined and across estimation methods. Mean CFI bias ranged from -1.35 (*SD* = 65.7) for the weighted least squares estimator in moderately nonnormal data conditions when the overall sample size equaled 300 to .049 (*SD* = .154) for the weighted least squares estimator in normal data conditions when the sample size equaled 300. In comparison to model one, the results were similar in each of the design conditions investigated. When the number of cluster size was fixed at 10, negative bias was detected for the maximum likelihood and generalized least squares estimators in severely nonnormal data conditions across all clusters. Note the least amount of bias related to the CFI was associated with the maximum likelihood estimator across all normal and nonormal data conditions investigated in the present study.

Table 61
Descriptive Results of CFI Bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	.036	.270	-.056	1.01	.049	.154
	N = 900 (30 x 30)	.024	.099	.027	.032	.031	.030
	N = 1500 (30 x 50)	-.008	.445	.008	.167	.016	.127
	N = 500 (50 x 10)	.004	.255	-.014	.527	.028	.209
	N = 1500 (50 x 30)	.025	.021	.024	.021	.026	.021
	N = 2500 (50 x 50)	-.069	2.13	.038	.179	.058	.080
	N = 1000 (100 x 10)	.005	.520	-.007	.539	.026	.319
	N = 3000 (100 x 30)	.026	.138	-.006	.532	.033	.048
	N = 5000 (100 x 50)	.047	.044	.046	.045	.048	.044
Moderately Nonnormal	N = 300 (30 x 10)	-.047	3.03	-.287	15.6	-1.35	65.7
	N = 900 (30 x 30)	-.058	1.38	-.079	1.66	.001	.604
	N = 1500 (30 x 50)	.009	.267	.014	.176	.022	.075
	N = 500 (50 x 10)	-.061	2.13	-.018	.622	-.033	2.56
	N = 1500 (50 x 30)	-.012	.386	-.021	.415	.010	.230
	N = 2500 (50 x 50)	.013	1.64	.009	2.62	.071	.048
	N = 1000 (100 x 10)	-.020	1.42	-.035	1.45	-.011	1.33
	N = 3000 (100 x 30)	.007	.392	.009	.323	.015	.478
	N = 5000 (100 x 50)	.022	.886	.035	.326	.046	.051
Severely Nonnormal	N = 300 (30 x 10)	-.044	2.38	-.319	15.9	.011	1.11
	N = 900 (30 x 30)	.002	.417	-.006	.453	.011	.306
	N = 1500 (30 x 50)	.009	.187	.003	.233	.013	.179
	N = 500 (50 x 10)	-.028	1.18	-.049	1.39	-.172	4.45
	N = 1500 (50 x 30)	.002	.249	-.016	.653	.013	.191
	N = 2500 (50 x 50)	.019	1.29	.045	.299	.061	.259
	N = 1000 (100 x 10)	-.036	1.44	-.054	1.92	.005	.779
	N = 3000 (100 x 30)	.007	.557	.014	.482	.022	.564
	N = 5000 (100 x 50)	.013	.354	-.205	11.7	.028	.372

The results reported in Table 62 indicate that bias among the root mean square residual (RMR) was similar across estimation methods in normal data conditions. As data increased in departure from normality, the generalized least squares estimator produced slightly more biased RMR results in moderately nonnormal data when compared to the maximum likelihood and weighted least squares estimators. Note the standard deviation associates with RMR bias among the weighted least squares estimates were greater in magnitude indicating that there was increased instability in the RMR bias associated with the weighted least squares estimator. The greatest amount of RMR bias was noted when the number of clusters and cluster size was 50 for the maximum likelihood estimator (*M* = -.513, *SD* = .391) in normal data conditions. Average RMR bias ranged from -.513 (SD = .391) with the maximum likelihood estimator in normal data to m = .313 (SD = .992) for the weighted least squares estimator in normal data conditions when the sample size equaled 5,000 (100 clusters with cluster size of 50).

Table 62
Descriptive Results of RMR Bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	-.176	.614	-.180	.649	-.165	.650
	N = 900 (30 x 30)	.025	.808	.123	.855	.058	.824
	N = 1500 (30 x 50)	-.266	.576	-.274	.585	-.248	.583
	N = 500 (50 x 10)	.274	.958	.178	.921	.238	.982
	N = 1500 (50 x 30)	-.201	.590	-.248	.586	-.243	.566
	N = 2500 (50 x 50)	-.491	.388	-.471	.389	-.469	.402
	N = 1000 (100 x 10)	-.188	.597	-.205	.619	-.175	.667
	N = 3000 (100 x 30)	-.513	.391	-.477	.383	-.495	.375
	N = 5000 (100 x 50)	.220	.974	.287	1.012	.313	.992
Moderately Nonnormal	N = 300 (30 x 10)	-.214	.604	-.217	.600	-.235	.597
	N = 900 (30 x 30)	-.038	.754	-.040	.734	-.058	.713
	N = 1500 (30 x 50)	-.234	.582	-.239	.584	-.238	.586
	N = 500 (50 x 10)	.170	.899	.217	.951	.210	.933
	N = 1500 (50 x 30)	-.285	.558	-.266	.569	-.269	.557
	N = 2500 (50 x 50)	-.436	.437	-.446	.419	-.442	.416
	N = 1000 (100 x 10)	-.165	.632	-.184	.626	-.204	.607
	N = 3000 (100 x 30)	-.503	.375	-.506	.375	-.512	.372
	N = 5000 (100 x 50)	.198	.963	.208	.916	.190	.933
Severely Nonnormal	N = 300 (30 x 10)	-.183	.637	-.208	.601	-.222	.606
	N = 900 (30 x 30)	-.046	.726	-.025	.736	-.052	.734
	N = 1500 (30 x 50)	-.261	.579	-.253	.579	-.281	.545
	N = 500 (50 x 10)	.217	.940	.212	.925	.236	.954
	N = 1500 (50 x 30)	-.278	.559	-.261	.559	-.282	.540
	N = 2500 (50 x 50)	-.445	.429	-.444	.427	-.453	.428
	N = 1000 (100 x 10)	-.164	.632	-.149	.654	-.183	.645
	N = 3000 (100 x 30)	-.495	.396	-.508	.372	-.495	.379
	N = 5000 (100 x 50)	.173	.914	.188	.924	.175	.935

Table 63 displays the descriptive results of AIC bias by data normality conditions, sample size and estimation method. The outcome revealed that bias among the maximum likelihood and generalized least squares estimators was similar in magnitude in normal and moderately nonnormal data conditions while the weighted least squares estimator produced negatively biased results in these same conditions. In severely nonnormal data conditions, the generalized least squares estimator produced slightly more biased results than the maximum likelihood estimator, especially in smaller sample sizes, while the weighted least squares estimator produced negatively biased results. Although similar results were reported for the weighted least squares estimator in each condition examined, AIC bias decreased with the weighted least squares estimator in severely nonnormal data conditions. Note the standard deviation associated with the weighted least squares estimator was almost one standard deviation less, in most instances, than the standard deviation associated with the maximum likelihood and generalized least squares estimators.

Compared to model one, AIC bias decreased in magnitude in model two. A plausible explanation could be attributed to the additional level-1 parameter added in model two. Mean AIC bias ranged from -.634 (*SD* = .491) for the weighted least squares estimator in normal data conditions when the number of clusters was 50 and cluster size was 30 to.123 (*SD* = 1.745) for the maximum likelihood estimator in severely nonnormal data when the sample size equaled 300 (i.e., number of clusters = 30 and cluster size =10).

Table 63
Descriptive Results of AIC bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	-.010	1.365	.046	1.587	-.521	.713
	N = 900 (30 x 30)	-.066	1.386	.092	1.547	-.635	.546
	N = 1500 (30 x 50)	.008	1.427	-.009	1.424	-.626	.510
	N = 500 (50 x 10)	.060	1.430	-.047	1.421	-.538	.648
	N = 1500 (50 x 30)	.078	1.503	.003	1.501	-.634	.491
	N = 2500 (50 x 50)	-.098	1.295	-.043	1.356	-.654	.486
	N = 1000 (100 x 10)	-.045	1.280	-.033	1.421	-.557	.675
	N = 3000 (100 x 30)	-.008	1.455	.076	1.532	-.635	.493
	N = 5000 (100 x 50)	-.081	1.362	-.022	1.363	-.633	.504
Moderately Nonnormal	N = 300 (30 x 10)	.016	1.438	.034	1.519	-.504	.804
	N = 900 (30 x 30)	.019	1.469	.017	1.433	-.604	.572
	N = 1500 (30 x 50)	-.005	1.398	.002	1.411	-.605	.562
	N = 500 (50 x 10)	-.053	1.358	.046	1.528	-.502	.710
	N = 1500 (50 x 30)	-.035	1.375	.032	1.459	-.597	.595
	N = 2500 (50 x 50)	.026	1.452	-.034	1.327	-.621	.535
	N = 1000 (100 x 10)	.029	1.436	-.003	1.425	-.550	.659
	N = 3000 (100 x 30)	-.009	1.417	-.007	1.394	-.600	.570
	N = 5000 (100 x 50)	-.018	1.475	-.033	1.311	-.629	.532
Severely Nonnormal	N = 300 (30 x 10)	.123	1.745	.067	1.554	-.312	1.084
	N = 900 (30 x 30)	.021	1.471	.063	1.509	-.489	.790
	N = 1500 (30 x 50)	.027	1.472	.047	1.483	-.567	.660
	N = 500 (50 x 10)	-.013	1.408	.009	1.555	-.336	1.024
	N = 1500 (50 x 30)	.037	1.498	.074	1.512	-.499	.770
	N = 2500 (50 x 50)	.023	1.496	.033	1.465	-.570	.681
	N = 1000 (100 x 10)	-.022	1.373	.040	1.477	-.371	.989
	N = 3000 (100 x 30)	.027	1.498	-.037	1.353	-.516	.749
	N = 5000 (100 x 50)	.011	1.436	.018	1.426	-.567	.719

Bivariate Results

Pearson product-moment correlations were calculated to determine how bias among the selected fit indices was impacted by the number of clusters, cluster size, correlation among variables, and data normality conditions in model two. Recall model two included two level-1 predictors, one level-2 predictor, and a cross-level interaction term. The results displayed in Table 64 show that statistically significant correlations ranged from r = -.061 between cluster size and the root mean square residual index (95% CI ranged from -.065 to -.056) to r = .292 between cluster size and the goodness of fit index (95% CI ranged from .287 to .296). The effect size was .085, indicating that cluster size explained approximately 9% of the variance in the goodness of fit index. Although the majority of the correlations among the study design conditions and fit indices were statistically significant, the effect sizes indicated that these conditions explained less than one percent of the variation in the outcome variable. Compared to model one, the most salient correlation noted in Table 61 was between cluster size and the goodness of fit index. A plausible explanation for the increased correlation between cluster size and the goodness of fit index could be attributed to the additional parameter in model two.

Table 64

Pearson Product-Moment Correlations among Fit Indices Bias and Selected Design Conditions

	1	2	3	4	5	6	7	8	9	10
GFI Bias (1)	1.00									
CFI Bias (2)	.012**	1.00								
Fit Bias (3)	-.639**	-.011**	1.00							
RMR Bias (4)	-.476**	-.008**	.774**	1.00						
AIC Bias (5)	-.613**	-.010**	.960**	.770**	1.00					
No. of Clusters (6)	.013**	.004	-.015**	.004	-.010**	1.00				
Cluster Size (7)	.292**	.006**	-.058**	-.061**	-.020**	-.004	1.00			
Correlation among Variables (8)	-.075**	.004	.033**	-.007**	.002	.000	.000	1.00		
Skew (9)	-.057**	.001	.040**	-.004	.041**	.001	.007**	.000	1.00	
Kurtosis (10)	-.045**	.000	.018**	-.006*	.019**	.000	.006*	.000	.396**	1.00

**.Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed). The sample size for each coefficient was 181,875.

Factorial Analysis Results of Fit Indices

Table 65 displays the factorial ANOVA results of the six study design conditions on GFI bias. Each of the main effects was statistically significant with the number of clusters and cluster size explaining the greatest amount of variation in the outcome variable. The number of clusters explained approximately 14% of the variance (partial η^2 = .139), while the clusters sizes was responsible for approximately 9% (partial η^2 = .084) of the variance in GFI bias. Regarding the interaction effects, the number of clusters by cluster size and the term comprised of the number of clusters by correlation among variables each explained approximately 3% of the variance in the criterion variable. Post hoc confidence intervals were calculated to determine how the number of clusters, and cluster size impacted GFI bias in model two. The results are displayed in Figure 20.

Table 65
Factorial ANOVA Results of the Study's Six Design Conditions Effect on GFI Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	.094	88	.001	998.727	< .001	.327[a]
Intercept	.237	1	.237	221903.698	< .001	.551
Est. Mthd	.001	2	.000	293.128	< .001	.003
Clstrs	.031	2	.016	14564.683	< .001	.139
Clstr. Size	.018	2	.009	8250.249	< .001	.084
Corr	.001	2	.001	501.665	< .001	.006
Skew	.000	3	.000	123.302	< .001	.002
Kurt	4.63E-005	3	1.54E-005	14.469	< .001	.000
Est. Mthd. * Clstrs	.000	4	8.21E-005	76.894	< .001	.002
Est. Mthd. * Clstr. Size	.001	4	.000	169.500	< .001	.004
Est. Mthd. * Corr	1.88E-006	4	4.69E-007	.439	.780	.000
Est. Mthd. * Skew	.001	6	.000	100.856	< .001	.003
Est. Mthd. * Kurt	3.96E-005	6	6.59E-006	6.178	< .001	.000
Clstrs * Clstr. Size	.005	4	.001	1282.551	< .001	.028
Clstrs * Corr	.006	4	.002	1407.530	< .001	.030
Clstrs * Skew	8.88E-005	6	1.48E-005	13.870	< .001	.000
Clstrs * Kurt	1.46E-005	6	2.44E-006	2.285	.033	.000
Clstr. Size * Corr	.002	4	.000	375.852	< .001	.008
Clstr. Size * Skew	.000	6	2.60E-005	24.390	< .001	.001
Clstr. Size * Kurt	2.09E-005	6	3.49E-006	3.271	.003	.000
Corr * Skew	1.07E-005	6	1.78E-006	1.669	.124	.000
Corr * Kurt	3.61E-006	6	6.02E-007	.564	.759	.000
Skew * Kurt	8.64E-006	2	4.32E-006	4.048	.017	.000
Error	.193	181036	1.07E-006			
Total	.584	181125				
Corrected Total	.287	181124				

a. R Squared = .327 (Adjusted R Squared = .326)

Note. The model eta squared for the full factorial 6-way model was .363 with df=696. The model eta squared for only the main and the two-way interaction effects was .327 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .036 (i.e., .363-.327).

Figure 20 displays the results of GFI bias by the number of clusters and cluster size by estimation method. The results revealed that in small sample size, especially when cluster size was 10 in both 30 and 50 clusters, the resulting GFI was biased downward with the most negative bias associated with the weighted least square estimator in 30 clusters with a cluster size of 10. Although the magnitude of the negative bias decreased as sample size increased, similar patterns were noted for 30 clusters with a cluster membership of 30 and 50 clusters with a cluster size of 10. GFI bias was negligible when the sample size increased to 1,000 and above.

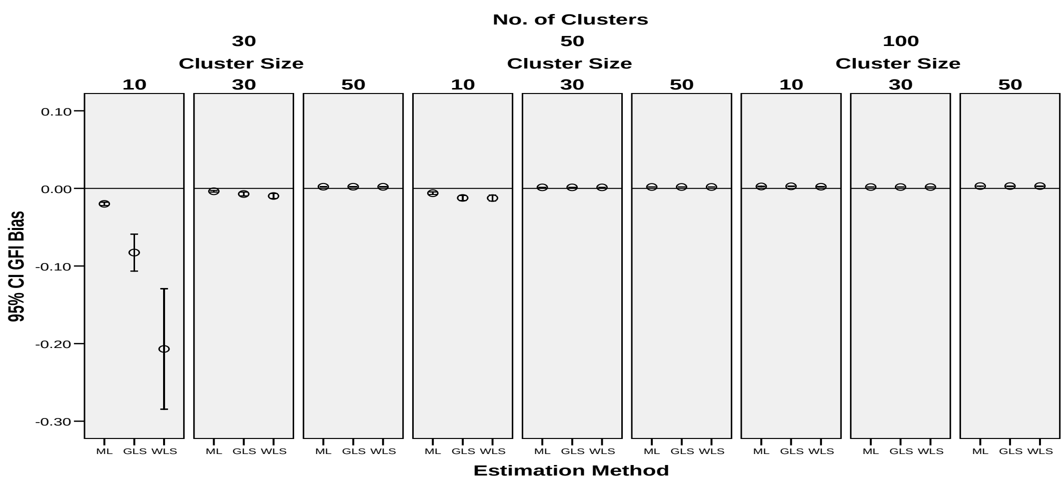

Figure 20. Confidence intervals examining GFI bias by estimation method, the number of clusters, and cluster size.

Regarding the impact of the study design conditions on Fit bias displayed in Table 66, with the exception of kurtosis, each of the main effects was statistically significant. The estimation method explained approximately two percent of the variation in the outcome variable, while cluster size was responsible for almost one percent of variation in fit bias. Remaining statistically significant main effects explained less than one percent of the variation in fit bias based on partial eta squared. As for the interaction effects, the terms including the number of clusters by skew, the number of clusters by kurtosis, and cluster size by kurtosis were not statistically significant while the number of clusters by cluster size explained one percent of the variation in the outcome variable, further indicating that fit bias was impacted by sample size. The post hoc analysis revealed that the maximum likelihood and generalized least squares estimators produced similar results, while the weighted least squares estimator produced negatively biased results. The results are displayed in Figure 21.

Table 66

Factorial ANOVA Results of the Study's Six Design Conditions Effect on Fit Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	17906.4	88	203.48	135.6	< .001	.062[a]
Intercept	6239.748	1	6239.7	4158	< .001	.022
Est. Mthd	6176.703	2	3088.4	2058	< .001	.022
Clstrs	162.326	2	81.163	54.09	< .001	.001
Clstr. Size	1022.254	2	511.13	340.6	< .001	.004
Corr	280.945	2	140.47	93.61	< .001	.001
Skew	325.809	3	108.60	72.37	< .001	.001
Kurt	5.286	3	1.762	1.174	.318	.000
Est. Mthd. * Clstrs	19.758	4	4.940	3.292	.010	.000
Est. Mthd. * Clstr. Size	31.060	4	7.765	5.174	< .001	.000
Est. Mthd. * Corr	34.524	4	8.631	5.752	< .001	.000
Est. Mthd. * Skew	540.733	6	90.122	60.06	< .001	.002
Est. Mthd. * Kurt	48.858	6	8.143	5.426	< .001	.000
Clstrs * Clstr. Size	2726.770	4	681.69	454.3	< .001	.010
Clstrs * Corr	672.367	4	168.09	112.0	< .001	.002
Clstrs * Skew	5.818	6	.970	.646	.693	.000
Clstrs * Kurt	2.769	6	.462	.308	.933	.000
Clstr. Size * Corr	655.941	4	163.99	109.3	< .001	.002
Clstr. Size * Skew	23.918	6	3.986	2.656	.014	.000
Clstr. Size * Kurt	7.066	6	1.178	.785	.582	.000
Corr * Skew	19.625	6	3.271	2.180	.042	.000
Corr * Kurt	5.548	6	.925	.616	.718	.000
Skew * Kurt	20.111	2	10.055	6.701	.001	.000
Error	271669	181036	1.501			
Total	298172	181125				
Corrected Total	289575	181124				

a. R Squared = .062 (Adjusted R Squared = .061)

Note. The model eta squared for the full factorial 6-way model was .072 with df=696. The model eta squared for only the main and the two-way interaction effects was .062 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .010 (i.e., .072-.062).

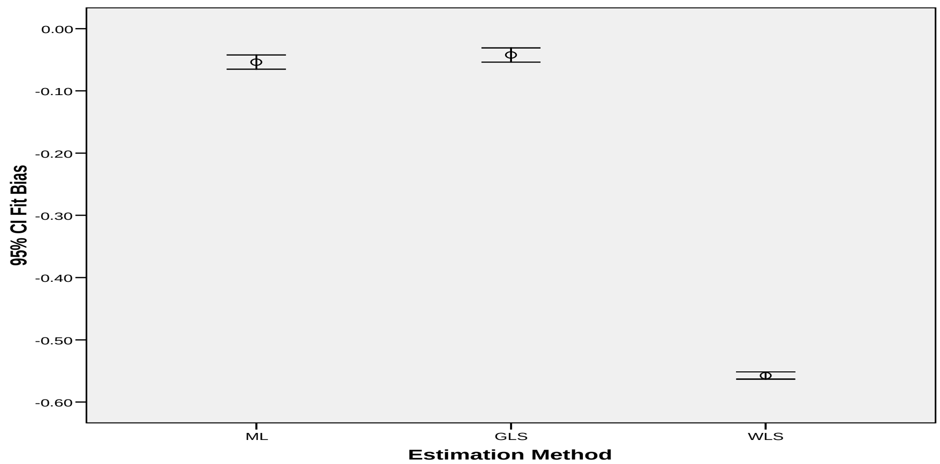

Figure 21. Confidence intervals examining fit bias by estimation method

Table 67 displays the factorial ANOVA results of the study design conditions on CFI bias. Interestingly, kurtosis was the only statistically significant main effect. However, when examining partial eta squared, kurtosis explained less than one percent of the variation in CFI bias. The findings contrast sharply with the results from model one. In model one, each of the main effects was statistically significant but, similar in model two, the resulting partial eta square associated with the statistically significant effects indicated that each explained less than one percent of the variance in the outcome variable.

The number of clusters and cluster size by kurtosis were the only statistically significant interaction terms, but were responsible for less than one percent of the variance in CFI bias. In comparison to model two, each of the interaction terms was statistically significant in model one. However, each statistically significant interaction effect explained less than one percent of the variance in the dependent variable. Figure 22 displays the post hoc confidence intervals examining the statistically significant study design conditions on CFI bias.

Table 67
Factorial ANOVA Results of the Study's Six Design Conditions Effect on CFI Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	9571.019	88	108.76	1.350	.016	.001[a]
Intercept	151.735	1	151.73	1.884	.170	.000
Est. Mthd	30.153	2	15.076	.187	.829	.000
Clstrs	98.338	2	49.169	.611	.543	.000
Clstr. Size	221.949	2	110.97	1.378	.252	.000
Corr	42.665	2	21.333	.265	.767	.000
Skew	279.016	3	93.005	1.155	.325	.000
Kurt	652.274	3	217.42	2.700	.044	.000
Est. Mthd. * Clstrs	232.079	4	58.020	.720	.578	.000
Est. Mthd. * Clstr. Size	313.117	4	78.279	.972	.421	.000
Est. Mthd. * Corr	337.339	4	84.335	1.047	.381	.000
Est. Mthd. * Skew	216.855	6	36.142	.449	.846	.000
Est. Mthd. * Kurt	567.816	6	94.636	1.175	.316	.000
Clstrs * Clstr. Size	613.660	4	153.41	1.905	.107	.000
Clstrs * Corr	666.592	4	166.65	2.069	.082	.000
Clstrs * Skew	514.920	6	85.820	1.066	.381	.000
Clstrs * Kurt	1080.026	6	180.00	2.235	.037	.000
Clstr. Size * Corr	711.185	4	177.80	2.208	.065	.000
Clstr. Size * Skew	741.707	6	123.62	1.535	.162	.000
Clstr. Size * Kurt	1364.345	6	227.39	2.823	.010	.000
Corr * Skew	352.205	6	58.701	.729	.626	.000
Corr * Kurt	638.689	6	106.45	1.322	.243	.000
Skew * Kurt	360.780	2	180.39	2.240	.106	.000
Error	13949058	173199	80.538			
Total	13958854	173288				
Corrected Total	13958629	173287				

a. R Squared = .001 (Adjusted R Squared = .000)

Note. The model eta squared for the full factorial 6-way model was .004 with df=696. The model eta squared for only the main and the two-way interaction effects was .001 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .003 (i.e., .004-.001).

The results displayed in Figure 22 revealed that in small sample size, especially when cluster size was held at 10, increasing values of kurtosis negatively impacted CFI bias. In other words, as kurtosis increased, from one to seven, negative CFI bias increased. As sample size increased above 1,000, CFI bias was negligible regardless of the degree of kurtosis.

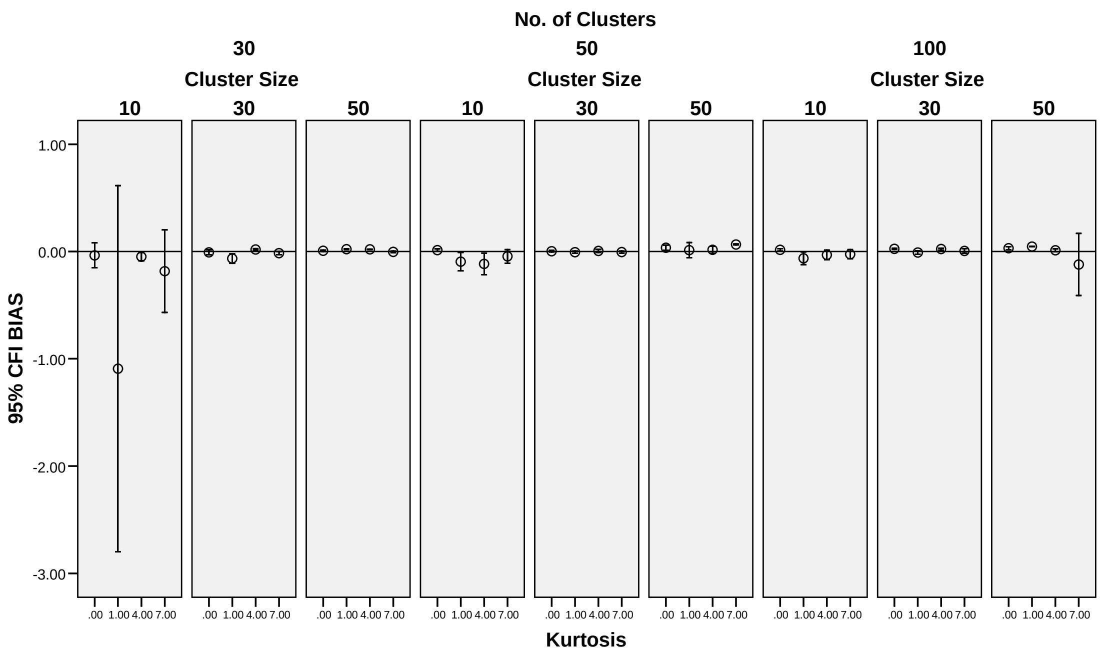

Figure 22. Confidence intervals examining CFI bias by the number of clusters, cluster size, and varying degrees of kurtosis.

Table 68 presents the results of investigating the study design conditions on RMR bias. Among the design conditions main effects, the number of clusters was not a statistically significant factor while the remaining conditions as main effects shared a statistically significant association with the outcome variable. Cluster size appeared to be the most salient main effect as determined by partial eta squared, which revealed that cluster size explained approximately one percent of the variation in the outcome variable while the remaining main effects were responsible for less than one percent of the variation in RMR bias. When the estimation method was included in the analysis as a two-way interaction term that included the estimation method by the remaining design conditions, the results were not statistically significant. However, the number of clusters by cluster size explained approximately 10% of the variance in the outcome variable (partial $\eta^2 = .098$), indicating that sample size was a salient factor impacting RMR bias. The remaining interaction terms were statistically significant but explained less than one percent of the variation in the dependent variable. The results from model two differ from the results obtained from model one. In model one, the interaction terms that included the estimation method were statistically significant while in model two, these same terms were not statistically significant. As for the remaining effects, both models produced similar results.

Table 68
Factorial ANOVA Results of the Study's Six Design Conditions Effect on RMR Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	10638	88	120.89	270.9	< .001	.116[a]
Intercept	4059.2	1	4059.2	9095	< .001	.048
Est. Mthd	3.753	2	1.877	4.204	.015	.000
Clstrs	.879	2	.439	.984	.374	.000
Clstr. Size	988.418	2	494.21	1107	< .001	.012
Corr	16.192	2	8.096	18.14	< .001	.000
Skew	5.276	3	1.759	3.941	< .001	.000
Kurt	11.385	3	3.795	8.503	< .001	.000
Est. Mthd. * Clstrs	2.353	4	.588	1.318	.260	.000
Est. Mthd. * Clstr. Size	.919	4	.230	.515	.725	.000
Est. Mthd. * Corr	1.276	4	.319	.715	.582	.000
Est. Mthd. * Skew	3.124	6	.521	1.166	.321	.000
Est. Mthd. * Kurt	2.706	6	.451	1.010	.416	.000
Clstrs * Clstr. Size	8821.3	4	2205.3	4941	< .001	.098
Clstrs * Corr	95.231	4	23.808	53.34	< .001	.001
Clstrs * Skew	28.627	6	4.771	10.69	< .001	.000
Clstrs * Kurt	9.897	6	1.650	3.696	.001	.000
Clstr. Size * Corr	207.476	4	51.869	116.2	< .001	.003
Clstr. Size * Skew	28.416	6	4.736	10.61	< .001	.000
Clstr. Size * Kurt	20.889	6	3.482	7.801	< .001	.000
Corr * Skew	32.069	6	5.345	11.98	< .001	.000
Corr * Kurt	20.222	6	3.370	7.551	< .001	.000
Skew * Kurt	4.181	2	2.091	4.684	.009	.000
Error	80798	181036	.446			
Total	96405	181125				
Corrected Total	91436	181124				

a. R Squared = .116 (Adjusted R Squared = .116)

<u>Note</u>. The model eta squared for the full factorial 6-way model was .127 with <u>df</u>=696. The model eta squared for only the main and the two-way interaction effects was .116 with <u>df</u>=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .011 (i.e., .127-.116).

The results displayed in Figure 23 revealed that of the six study design conditions examined, the number of clusters and cluster size had the greatest impact on RMR bias. As the correlation among variables increased from r= .30 to r = .70, the variation in RMR bias increased in small and moderately small sample sizes. As sample size increased, RMR bias decreased across estimation methods.

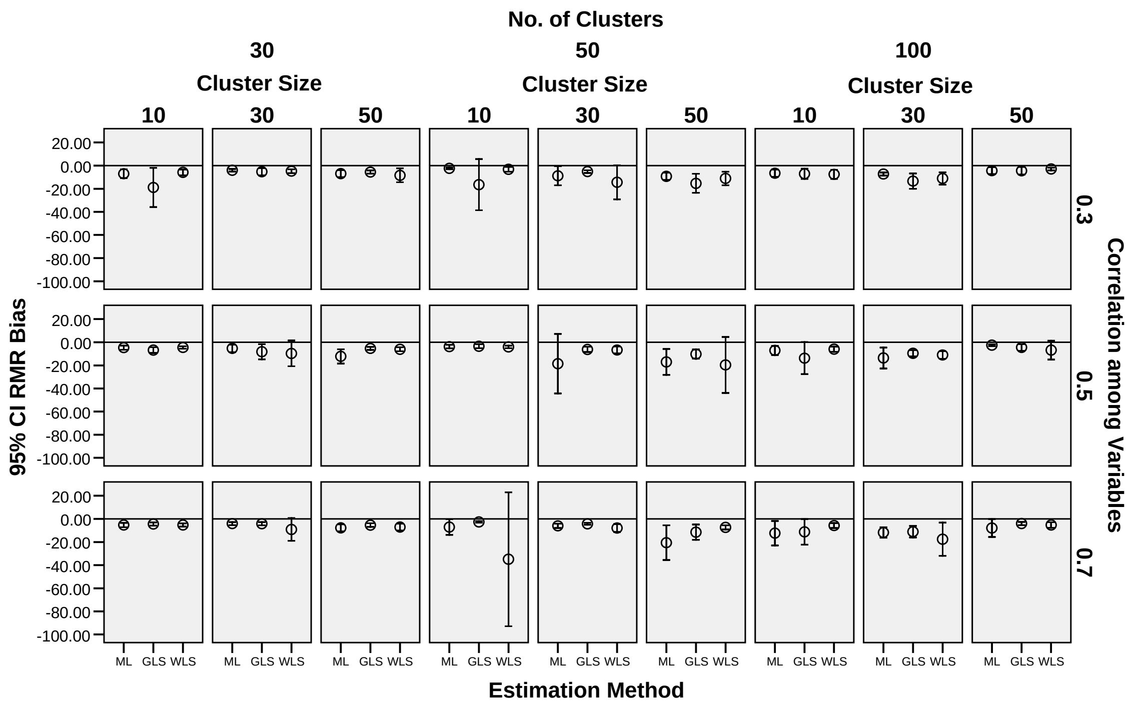

Figure 23. Confidence intervals examining RMR bias by estimation method, the number of clusters, cluster size, and the correlation among variables

The influence of the study design conditions on AIC bias are displayed in Table 69. Each of the main effects with the exception of kurtosis was statistically significant. The estimation method returned the largest partial eta squared (partial η^2 =.024), while the remaining statistically significant main effects explained less than one percent of the variation in the outcome variable. As for the interaction effects that included estimation method, the terms comprised of estimation method by cluster size, estimation method by skew, and estimation method by kurtosis were statistically significant. However, each explained less than one percent of the variance in the outcome variable. In comparison to model one, the estimation method as a main effect for model two was not a statistically significant factor.

Regarding interaction terms, the term that included the estimation method by correlation among variables was statistically significant in model one, while the same term was not a statistically significant effect in model two. The resulting post hoc analysis of AIC bias by estimation method is displayed in Figure 24.

Table 69

Factorial ANOVA Results of the Study's Six Design Conditions Effect on AIC Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	13812.6	88	156.96	99.62	< .001	.046[a]
Intercept	3644.17	1	3644.2	2313	< .001	.013
Est. Mthd	7123.38	2	3561.7	2260	< .001	.024
Clstrs	20.955	2	10.478	6.650	.001	.000
Clstr. Size	96.120	2	48.060	30.50	< .001	.000
Corr	3.301	2	1.651	1.048	.351	.000
Skew	347.708	3	115.90	73.56	< .001	.001
Kurt	4.451	3	1.484	.942	.419	.000
Est. Mthd. * Clstrs	4.055	4	1.014	.643	.631	.000
Est. Mthd. * Clstr. Size	154.688	4	38.672	24.54	< .001	.001
Est. Mthd. * Corr	2.571	4	.643	.408	.803	.000
Est. Mthd. * Skew	624.374	6	104.06	66.05	< .001	.002
Est. Mthd. * Kurt	62.043	6	10.340	6.563	< .001	.000
Clstrs * Clstr. Size	9.302	4	2.325	1.476	.206	.000
Clstrs * Corr	8.092	4	2.023	1.284	.274	.000
Clstrs * Skew	4.016	6	.669	.425	.863	.000
Clstrs * Kurt	4.269	6	.711	.452	.844	.000
Clstr. Size * Corr	14.598	4	3.649	2.316	.055	.000
Clstr. Size * Skew	17.995	6	2.999	1.904	.076	.000
Clstr. Size * Kurt	7.710	6	1.285	.816	.558	.000
Corr * Skew	15.422	6	2.570	1.631	.134	.000
Corr * Kurt	5.960	6	.993	.630	.706	.000
Skew * Kurt	21.145	2	10.573	6.710	.001	.000
Error	285244	181036	1.576			
Total	304232	181125				
Corrected Total	299057	181124				

a. R Squared = .046 (Adjusted R Squared = .046)

Note. The model eta squared for the full factorial 6-way model was .049 with df=696. The model eta squared for only the main and the two-way interaction effects was .046 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .003 (i.e., .049-.046).

The resulting post hoc confidence intervals examining AIC bias by estimation method displayed in Figure 24 revealed that the maximum likelihood and generalized least squares estimators produced similar results, while the weighted least squares estimator produce negatively biased results. Similar results were noted when examining the impact of the estimation method on AIC bias in each of the remaining study design conditions.

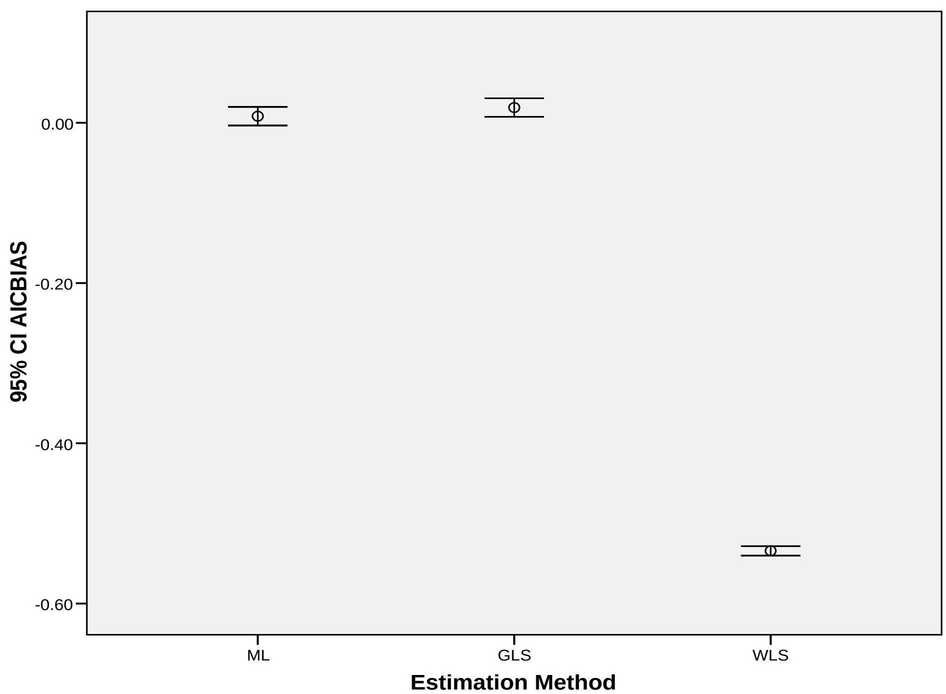

Figure 24. Confidence intervals examining AIC Bias by estimation method.

Summary of Model Two

Similar to the results from model one, the addition of the level one parameter in model two indicated that of the six conditions examined, sample size appeared to have the greatest impact on the goodness of fit index, while the estimation method impacted the fit and AIC indices (partial $\eta^2 = \sim.02$). Among the estimation methods, there was negligible bias associated with the maximum likelihood and generalized least squares estimators, while severe negative bias was found for both the fit and AIC indices among the weighted least squares estimation method. Regarding CFI bias, kurtosis as a main effect and the number of clusters by kurtosis and cluster size by kurtosis as interaction terms had a statistically significant impact on CFI bias. However, each of the statistically significant effects explained less than one percent of the bias associated with CFI.

The results of the six study conditions regarding bias among the root mean square residual revealed that the number of clusters by cluster size had the greatest impact on RMR bias (partial η^2 =.098). As the correlation among variables increased from *r* = .30 to *r* = .70, the variation in RMR bias increased in small and moderately small sample sizes. As sample size increased, RMR bias decreased across estimation methods.

Model Three

Model three was comprised of two level-1, two level-2 predictor variables and two cross-level interaction terms. The initial cross-level interaction term included the initial level-1 and the initial level-2 variables, while the subsequent interaction term included the subsequent level-1 and level-2 predictor variables. The model is displayed below.

Level-1 equation $Y_{ij} = \beta_0 + \beta_1 X1_{ij} + \beta_2\ X4_{ij} + \varepsilon_{ij}$

Level-2 equation $\beta_0 = \gamma 00 + U0_j$

$\beta_1 = \gamma 10 + \gamma_{11} XC2BAR$

$\beta_2 = \gamma 20 + \gamma 21\ XC5BAR$

Combined equation $Y = \gamma_{00} + \gamma_0 X1 + \gamma_{11}\ X2BAR*X1 + \gamma_{20}\ x_4 +$

$\gamma_{21}*X5BAR*X_4\ U_0 + \varepsilon$

Descriptive Measures of Bias among Selected Fit Indices

The descriptive results of GFI bias presented in Table 70 for model three showed that bias was negligible across estimation methods in each of the data normality and sample sizes investigated. Comparing the results to those reported for model one and two, there was greater variation in GFI bias, especially when the number of clusters was 30 and cluster size was held at 10 in each data normality condition and across estimation methods. Note in moderately nonnormal data conditions, that GFI bias associated with the generalized and weighted least squares estimators was twice as large as the GFI bias associated with the maximum likelihood estimator when the number of clusters was 30 and cluster size equaled 10. As noted earlier, the GIF bias and variation in GFI bias decreased as sample size increased.

Regarding the fit index, the greatest amount of variation in fit index bias was in small sample sizes especially when the number of clusters equaled 30 in each data normality condition investigated. As sample size increased, bias associated with the fit index decreased across estimation methods. When cluster size was greater than the number of clusters, fit index was somewhat erratic as indicated by the mean and standard deviation of fit bias when the number of clusters was 30 and cluster size was 50. In this scenario, the weighted least squares estimator produced the least amount of bias in each of the data normality conditions investigated. The results of the descriptive analysis(ie.e, means and standadrd deviation among the selected six study design conditions and fit index bias are reported in Table 71 below.

Table 70
Descriptive Results of GFI Bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluater Size	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	.061	.274	.035	.457	-.014	.136
	N = 900 (30 x 30)	-.006	.047	-.010	.036	-.014	.021
	N = 1500 (30 x 50)	-.003	.041	-.005	.024	-.007	.025
	N = 500 (50 x 10)	-.016	.150	-.029	.091	-.026	.166
	N = 1500 (50 x 30)	-.011	.031	-.012	.026	-.016	.017
	N = 2500 (50 x 50)	.000	.006	.000	.006	-.001	.003
	N = 1000 (100 x 10)	-.022	.043	-.024	.023	-.024	.028
	N = 3000 (100 x 30)	.001	.011	.001	.007	.000	.006
	N = 5000 (100 x 50)	-.001	.001	-.001	.001	-.001	.001
Moderately Nonnormal	N = 300 (30 x 10)	.162	.383	2.208	14.704	1.368	12.169
	N = 900 (30 x 30)	.037	.209	.036	.369	.015	.251
	N = 1500 (30 x 50)	.028	.158	.020	.188	.012	.144
	N = 500 (50 x 10)	.058	.282	.434	4.183	.497	5.219
	N = 1500 (50 x 30)	-.014	.027	-.014	.018	-.016	.015
	N = 2500 (50 x 50)	-.001	.004	.000	.004	-.001	.002
	N = 1000 (100 x 10)	-.028	.018	-.028	.016	-.029	.015
	N = 3000 (100 x 30)	.000	.009	-.001	.008	-.001	.005
	N = 5000 (100 x 50)	.000	.002	.000	.002	.000	.001
Severely Nonnormal	N = 300 (30 x 10)	.022	.226	.017	.389	.083	1.021
	N = 900 (30 x 30)	.007	.129	-.005	.078	-.004	.101
	N = 1500 (30 x 50)	.001	.062	-.002	.040	-.002	.050
	N = 500 (50 x 10)	-.023	.112	-.028	.083	-.030	.102
	N = 1500 (50 x 30)	-.006	.079	-.011	.050	-.010	.054
	N = 2500 (50 x 50)	.001	.010	.001	.010	.000	.007
	N = 1000 (100 x 10)	-.007	.117	-.015	.072	-.017	.068
	N = 3000 (100 x 30)	-.001	.004	-.001	.004	-.002	.004
	N = 5000 (100 x 50)	.000	.001	.000	.001	.000	.001

Table 71

Descriptive Results of Fit Bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Cluisters x Cluste Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	-.672	18.730	.126	4.174	.511	1.881
	N = 900 (30 x 30)	.538	1.519	.648	1.181	.767	.643
	N = 1500 (30 x 50)	-2.655	23.282	-1.408	9.989	-.946	10.624
	N = 500 (50 x 10)	.615	5.339	.828	.890	.769	1.950
	N = 1500 (50 x 30)	.471	1.693	.565	1.095	.782	.486
	N = 2500 (50 x 50)	-.811	4.628	-.609	4.212	.137	2.462
	N = 1000 (100 x 10)	.508	5.033	.747	.860	.725	1.123
	N = 3000 (100 x 30)	-.699	4.968	-.517	3.501	.038	2.760
	N = 5000 (100 x 50)	.571	1.156	.538	1.221	.820	.657
Moderately Nonnormal	N = 300 (30 x 10)	-2.126	70.113	-16.724	119.352	-10.095	94.441
	N = 900 (30 x 30)	-2.190	57.470	-1.130	13.007	-.146	7.332
	N = 1500 (30 x 50)	-11.797	177.665	-10.247	90.216	-6.850	60.378
	N = 500 (50 x 10)	-.255	19.533	-3.684	40.756	-4.201	45.611
	N = 1500 (50 x 30)	.712	1.014	.723	.879	.831	.693
	N = 2500 (50 x 50)	.159	1.645	.089	2.029	.651	.904
	N = 1000 (100 x 10)	.904	.511	.913	.338	.952	.195
	N = 3000 (100 x 30)	.080	4.224	.168	3.565	.543	2.110
	N = 5000 (100 x 50)	-.077	2.300	-.064	2.133	.594	.872
Severely Nonnormal	N = 300 (30 x 10)	-.362	35.089	.522	1.812	.413	2.930
	N = 900 (30 x 30)	-.498	25.674	.303	3.808	.474	3.609
	N = 1500 (30 x 50)	-6.171	114.287	-2.581	18.43	-1.987	18.620
	N = 500 (50 x 10)	.079	31.489	.835	.758	.895	.612
	N = 1500 (50 x 30)	-3.313	125.115	-.215	8.485	.270	4.427
	N = 2500 (50 x 50)	-.907	5.266	-1.005	6.743	.260	2.217
	N = 1000 (100 x 10)	-.762	41.081	.462	2.898	.674	1.873
	N = 3000 (100 x 30)	.535	1.173	.550	1.202	.803	.609
	N = 5000 (100 x 50)	.275	1.593	.302	1.555	.766	.440

Similar to models one and two, the results reported in Table 72 indicate that CFI bias was similar across estimation methods with the weighted and generalized least squares estimators producing slightly more biased results when the number of clusters equaled 30 and cluster size was held at 10, while the maximum likelihood estimator produced negligible bias. The largest amount of bias was noted when cluster size was 10 and the number of clusters was 30 and 50 respectively. As sample size increased from 300, each of the estimation methods produced similar results with little to no bias reported. Note in small sample sizes (i.e., 300 and 900), the generalized least squares estimator produced the greatest amount of CFI bias with similar results reported for the weighted least squares estimator. Overall, however, the least amount of CFI bias was detected for the maximum likelihood estimator.

Table 72
Descriptive Results of CFI Bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	.060	.344	.181	1.079	.053	.316
	N = 900 (30 x 30)	.005	.069	.035	.185	.017	.113
	N = 1500 (30 x 50)	-.004	.031	.008	.063	.002	.060
	N = 500 (50 x 10)	.005	.434	.004	.203	.004	.377
	N = 1500 (50 x 30)	-.014	.032	-.002	.060	-.012	.039
	N = 2500 (50 x 50)	.017	.072	.023	.087	.010	.051
	N = 1000 (100 x 10)	-.016	.039	.002	.075	-.005	.065
	N = 3000 (100 x 30)	-.001	.011	.007	.026	.003	.028
	N = 5000 (100 x 50)	.000	.007	.002	.014	.000	.006
Moderately Nonnormal	N = 300 (30 x 10)	.203	1.326	5.394	34.503	3.116	26.124
	N = 900 (30 x 30)	.049	.561	.126	.863	.076	.626
	N = 1500 (30 x 50)	.032	.204	.080	.494	.050	.318
	N = 500 (50 x 10)	.078	.724	1.194	10.366	1.054	8.996
	N = 1500 (50 x 30)	-.017	.033	-.005	.054	-.014	.034
	N = 2500 (50 x 50)	-.003	.008	.002	.018	-.002	.009
	N = 1000 (100 x 10)	-.016	.032	-.010	.045	-.016	.032
	N = 3000 (100 x 30)	-.001	.020	.003	.027	-.001	.019
	N = 5000 (100 x 50)	.002	.012	.004	.014	.001	.006
Severely Nonnormal	N = 300 (30 x 10)	.073	1.389	.221	1.283	.240	1.765
	N = 900 (30 x 30)	.015	.239	.036	.251	.036	.310
	N = 1500 (30 x 50)	.006	.098	.024	.111	.016	.120
	N = 500 (50 x 10)	.015	1.729	.011	.205	.000	.220
	N = 1500 (50 x 30)	-.001	.403	.005	.146	.002	.132
	N = 2500 (50 x 50)	.001	.019	.010	.044	.004	.023
	N = 1000 (100 x 10)	.010	.401	.021	.173	.012	.158
	N = 3000 (100 x 30)	-.004	.007	.000	.015	-.002	.012
	N = 5000 (100 x 50)	.001	.012	.002	.011	.001	.005

Similar to models one and two, the results displayed in Table 73 show that bias among the root mean square residual was similar across estimation methods in normal data conditions. As data increased in departure from normality, the weighted least squares estimator produced less biased RMR results in severely nonnormal data when compared to the maximum likelihood and generalized least squares estimators. Note the standard deviation associated with RMR bias among the weighted least squares estimates were greater in magnitude indicating that there was increased instability in the RMR bias associated with the weighted least squares estimator. The greatest amount of RMR bias was noted when cluster size was greater than the number of clusters (i.e., when the number of clusters equaled 30 and cluster size was held at 50) with negative RMR bias reported across each of the estimation methods investigated in the present study.

Table 73
Descriptive Results of RMR Bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	.106	.616	.000	.658	-.035	.689
	N = 900 (30 x 30)	-.120	.672	-.113	.650	-.151	.691
	N = 1500 (30 x 50)	-.252	.985	-.247	.983	-.272	.964
	N = 500 (50 x 10)	-.090	.760	-.160	.879	-.176	1.061
	N = 1500 (50 x 30)	-.060	.635	-.051	.620	-.107	.648
	N = 2500 (50 x 50)	.129	.517	.170	.481	.110	.530
	N = 1000 (100 x 10)	.068	.571	.051	.576	.059	.565
	N = 3000 (100 x 30)	.005	.565	.052	.569	.001	.568
	N = 5000 (100 x 50)	.000	.573	.004	.590	.010	.581
Moderately Nonnormal	N = 300 (30 x 10)	.204	.680	-.104	.952	-.279	1.220
	N = 900 (30 x 30)	-.006	.694	-.075	.713	-.127	.774
	N = 1500 (30 x 50)	-.225	.942	-.304	1.008	-.350	1.071
	N = 500 (50 x 10)	.070	.696	-.063	.776	-.220	1.358
	N = 1500 (50 x 30)	.006	.601	-.002	.604	-.007	.622
	N = 2500 (50 x 50)	.085	.560	.100	.569	.094	.578
	N = 1000 (100 x 10)	.080	.567	.079	.570	.077	.557
	N = 3000 (100 x 30)	.032	.594	.024	.580	.028	.584
	N = 5000 (100 x 50)	-.008	.575	-.032	.579	-.014	.570
Severel Nonnormal	N = 300 (30 x 10)	.078	.628	.004	.622	-.074	1.051
	N = 900 (30 x 30)	-.035	.653	-.046	.642	-.118	.749
	N = 1500 (30 x 50)	-.255	.993	-.252	1.013	-.344	1.157
	N = 500 (50 x 10)	.039	.639	.007	.640	-.001	.649
	N = 1500 (50 x 30)	-.015	.623	-.017	.622	-.072	.693
	N = 2500 (50 x 50)	.075	.554	.084	.544	.072	.565
	N = 1000 (100 x 10)	.037	.620	.018	.648	.018	.643
	N = 3000 (100 x 30)	.044	.579	.057	.552	.043	.564
	N = 5000 (100 x 50)	-.006	.573	.016	.574	.013	.571

As shown in Table 74, bias among AIC was varied in magnitude in small sample sizes especially when the number of clusters was 30 and cluster size ranged from 10 to 30. In these same conditions, the maximum likelihood estimator generally produced slightly less biased results in each of the data normality conditions examined. Mean AIC bias ranged from -4.00 (*SD* = .678) among the maximum likelihood estimator in moderately nonnormal data conditions in 30 clusters with a cluster size of 50 to 16.72 (*SD* =644.86) among the generalized least squares estimator in moderately nonnormal data in 30 clusters with a cluster size of 10. The number of clusters appeared to have the greatest impact on AIC bias. This was most evident in moderately nonnormal data conditions when the number of clusters was fixed at 30. In this scenario, negative bias was associated with each estimator with unstable results as indicated by the large standard deviations reported. Although bias was prevalent for each estimator, the weighted least squares estimator produced less biased results when the sample size was fixed at 900 and 1500.

Table 74
Descriptive Results of AIC Bias among Selected Design Conditions

		Estimation Method					
		ML		GLS		WLS	
Normality	Sample Size (No. Clusrters x Cluster Size)	M	SD	M	SD	M	SD
Normal	N = 300 (30 x 10)	-.578	18.199	.126	4.174	.511	1.881
	N = 900 (30 x 30)	.537	1.518	.647	1.181	.767	.643
	N = 1500 (30 x 50)	.174	4.639	.435	1.986	.581	2.115
	N = 500 (50 x 10)	.621	5.294	.828	.890	.769	1.950
	N = 1500 (50 x 30)	.471	1.693	.565	1.095	.782	.486
	N = 2500 (50 x 50)	-.811	4.628	-.609	4.212	.137	2.462
	N = 1000 (100 x 10)	.508	5.033	.747	.860	.725	1.123
	N = 3000 (100 x 30)	-.699	4.968	-.517	3.501	.038	2.760
	N = 5000 (100 x 50)	.571	1.156	.538	1.221	.820	.657
Moderately Nonnormal	N = 300 (30 x 10)	-1.653	64.601	-16.724	119.352	-10.095	94.441
	N = 900 (30 x 30)	-2.179	60.374	-1.215	13.50	-.207	7.709
	N = 1500 (30 x 50)	-4.000	49.253	-2.672	21.01	-1.650	16.158
	N = 500 (50 x 10)	-.168	18.845	-3.684	40.75	-4.201	45.611
	N = 1500 (50 x 30)	.712	1.014	.723	.879	.831	.693
	N = 2500 (50 x 50)	.159	1.645	.089	2.029	.651	.904
	N = 1000 (100 x 10)	.904	.511	.913	.338	.952	.195
	N = 3000 (100 x 30)	.080	4.224	.168	3.56	.543	2.110
	N = 5000 (100 x 50)	-.077	2.300	-.064	2.13	.594	.872
Severely Nonnormal	N = 300 (30 x 10)	-.308	34.386	.522	1.81	.413	2.930
	N = 900 (30 x 30)	-.487	25.523	.297	3.80	.471	3.610
	N = 1500 (30 x 50)	-.759	22.870	-.077	4.32	.230	3.948
	N = 500 (50 x 10)	.086	31.375	.835	.758	.895	.612
	N = 1500 (50 x 30)	-3.301	124.942	-.215	8.48	.270	4.427
	N = 2500 (50 x 50)	-.907	5.266	-1.005	6.74	.260	2.217
	N = 1000 (100 x 10)	-.750	40.933	.462	2.89	.674	1.873
	N = 3000 (100 x 30)	.535	1.173	.550	1.20	.803	.609
	N = 5000 (100 x 50)	.275	1.593	.302	1.55	.766	.440

Bivariate Results

To determine how bias among the selected fit indices was impacted by the number of clusters, cluster size, correlation among variables, and data normality conditions, Pearson product-moment correlations were calculated. The results displayed in Table 75 show that among the study design conditions investigated, statistically significant correlations ranged from r = -.034 between cluster size and the goodness of fit index bias (95% CI ranged from -.038 to -.029) to r = .075 between cluster size and AIC bias (95% CI ranged from .070 to .079). To interpret, as cluster size increased, bias associated with the goodness of fit index decreased. Regarding AIC bias, as cluster size increased, AIC bias increased. Although the majority of the correlations among the study design conditions and fit indices were statistically significant, squaring the correlation coefficients resulted in effect sizes that explained less than one percent of the variation in the outcome variable.

Table 75

Pearson Product-Moment Correlations among Fit Indices Bias and Selected Design Conditions

	1	2	3	4	5	6	7	8	9	10
GFI Bias (1)	1.00									
CFI Bias (2)	.989**	1.00								
Fit Bias (3)	-.464**	-.484**	1.00							
RMR Bias (4)	-.132**	-.128**	.142**	1.00						
AIC Bias (5)	-.632**	-.659**	.833**	.137**	1.00					
No. of Clusters (6)	-.028**	-.030**	.033**	.075**	.028**	1.00				
Cluster Size (7)	-.034**	-.038**	-.005*	-.033**	.013**	-.008**	1.00			
Correlation among Variables (8)	.017**	.017**	-.012**	-.010**	-.017**	-.008**	-.009**	1.00		
Skew (9)	-.027**	-.027**	.012**	.007**	.016**	.000	.000	.000	1.00	
Kurtosis (10)	-.009**	-.009**	.007**	-.015**	.007**	.000	.000	.000	-.186**	1.00

**.Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).The sample size for each coefficient was 181,875.

Factorial Analysis Results of Fit Indices

Table 76 displays the results of GFI bias by the six study design conditions for model three. Each of the design conditions as main and interaction effects was statistically significant. However, when examining the partial eta square results, each of the statistically significant effects explained less than one percent of the variance in the outcome variable. To provide further insight into the results, post hoc confidence intervals were calculated that included the number of clusters, cluster size, and estimation method. The results are displayed in Figure 25.

The results displayed in Figure 25 indicated that in small sample sizes, especially when cluster size was 10 and the number of clusters ranged from 30 to 50, GFI was negatively biased with the maximum likelihood estimator producing less biased results when compared to the generalized and weighted least squares estimators.

When comparing the generalized and weighted least squares estimators, both produced similar results as indicated by the overlapping confidence intervals when sample size was 300 and 500 respectively. As sample size increased, the results converged across estimation methods with little to no bias noted when the sample size increased beyond 500.

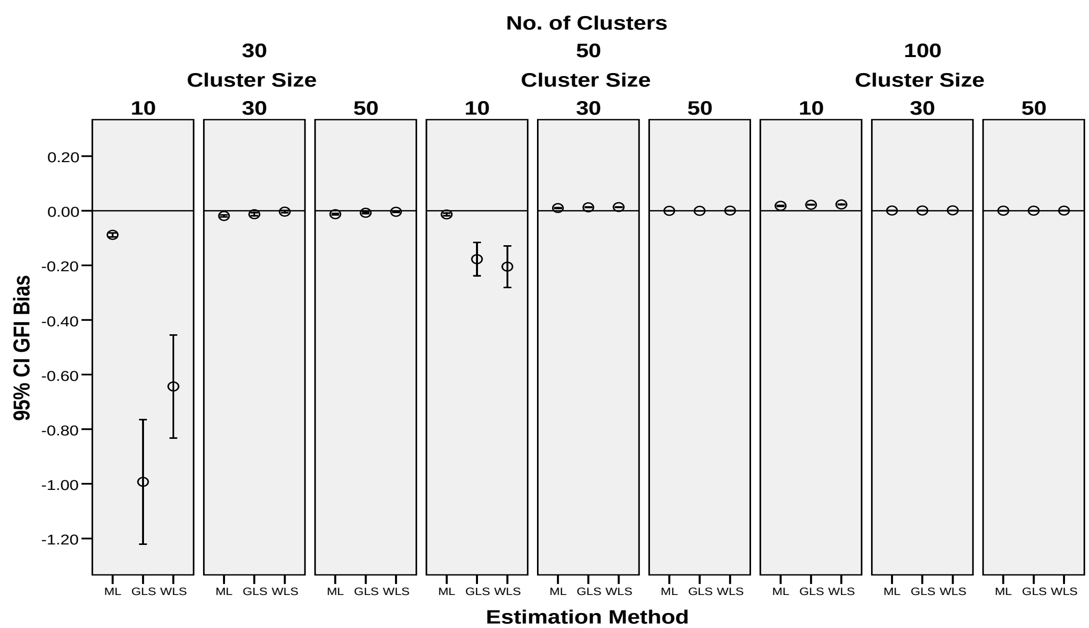

Figure 25. Confidence intervals examining GFI bias by estimation method, the number of clusters, and cluster size.

Table 76
Factorial ANOVA Results of the Study's Six Design Conditions Effect on GFI Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	37209.37	88	387.598	62.701	< .001	.027[a]
Intercept	1841.155	1	1841.16	297.840	< .001	.001
Est. Mthd	799.829	2	399.914	64.693	< .001	.001
Clstrs	2107.239	2	1053.62	170.442	< .001	.002
Clstr. Size	3913.513	2	1956.76	316.540	< .001	.003
Corr	1533.854	2	766.927	124.064	< .001	.001
Skew	19.865	3	6.622	1.071	< .001	.000
Kurt	1952.523	3	488.131	78.964	< .001	.001
Est. Mthd. * Clstrs	536.479	4	134.120	21.696	< .001	.000
Est. Mthd. * Clstr. Size	953.359	4	238.340	38.556	< .001	.001
Est. Mthd. * Corr	456.005	4	114.001	18.442	< .001	.000
Est. Mthd. * Skew	1.055	6	.176	.028	.897	.000
Est. Mthd. * Kurt	1091.365	8	136.421	22.068	< .001	.001
Clstrs * Clstr. Size	2640.113	4	660.028	106.771	< .001	.002
Clstrs * Corr	732.016	4	183.004	29.604	< .001	.001
Clstrs * Skew	26.547	6	4.425	.716	.637	.000
Clstrs * Kurt	2365.186	8	295.648	47.826	< .001	.002
Clstr. Size * Corr	1871.881	4	467.970	75.703	< .001	.001
Clstr. Size * Skew	12.116	6	2.019	.327	.923	.000
Clstr. Size * Kurt	3730.702	8	466.338	75.438	< .001	.003
Corr * Skew	9.639	6	1.607	.260	.955	.000
Corr * Kurt	4758.984	8	594.873	96.231	< .001	.004
Skew * Kurt	9.335	1	9.335	1.510	.219	.000
Error	1317956	213203	6.182			
Total	1356256	213300				
Corrected Total	1355166	213299				

a. R Squared = .027 (Adjusted R Squared = .027)

Note. The model eta squared for the full factorial 6-way model was .106 with df=696. The model eta squared for only the main and the two-way interaction effects was .027 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .079 (i.e., .106-.027).

Table 77 displays the results of the study’s design conditions on fit bias. The results revealed that the number of clusters, cluster size, correlation among variables, and kurtosis had a statistically significant impact on fit bias as main effects, while the estimation method was not statistically significant. Regarding interaction effects, estimation method by the skew was not a statistically significant effect while the remaining interaction terms shared a statistically significant relation with the outcome variable. Similar to GFI bias, the resulting partial eta square indicated that each of the statistically significant effects explained less than one percent of the variance in fit bias.

Post hoc confidence intervals were derived to examine fit bias by the number of clusters, cluster size, correlation among variables and kurtosis. The results are presented in Figure 26 below.

Table 77

Factorial ANOVA Results of the Study's Six Design Conditions Effect on Fit Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	5998086	88	62480.07	35.233	< .001	.016[a]
Intercept	326429.4	1	326429.4	184.079	< .001	.001
Est. Mthd	19059.726	2	9529.863	5.374	.005	.000
Clstrs	494448.2	2	247224.1	139.414	< .001	.001
Clstr. Size	173070.6	2	86535.29	48.799	< .001	.000
Corr	108813.1	2	54406.56	30.681	< .001	.000
Skew	208544.1	3	69514.70	39.200	< .001	.001
Kurt	339122.7	3	84780.68	47.809	< .001	.001
Est. Mthd. * Clstrs	15451.654	4	3862.913	2.178	< .001	.000
Est. Mthd. * Clstr. Size	67570.349	4	16892.59	9.526	< .001	.000
Est. Mthd. * Corr	72166.032	4	18041.51	10.174	< .001	.000
Est. Mthd. * Skew	19206.027	6	3201.004	1.805	.094	.000
Est. Mthd. * Kurt	55457.865	8	6932.233	3.909	< .001	.000
Clstrs * Clstr. Size	277561.6	4	69390.40	39.130	< .001	.001
Clstrs * Corr	230331.1	4	57582.77	32.472	< .001	.001
Clstrs * Skew	394010.0	6	65668.34	37.031	< .001	.001
Clstrs * Kurt	573266.2	8	71658.28	40.409	< .001	.002
Clstr. Size * Corr	727486.4	4	181871.6	102.560	< .001	.002
Clstr. Size * Skew	262444.5	6	43740.75	24.666	< .001	.001
Clstr. Size * Kurt	475963.8	8	59495.48	33.550	< .001	.001
Corr * Skew	326858.6	6	54476.43	30.720	< .001	.001
Corr * Kurt	746318.9	8	93289.87	52.608	< .001	.002
Skew * Kurt	28851.577	1	28851.58	16.270	< .001	.000
Error	3.8E+008	211982	1773.315			
Total	3.8E+008	212079				
Corrected Total	3.8E+008	212078				

Note. The model eta squared for the full factorial 6-way model was .053 with df=696. The model eta squared for only the main and the two-way interaction effects was .016 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .037 (i.e., .053-.016).

Figure 26 revealed that in normal data conditions when the correlation among variables was held at r = .70, that fit bias was negative when the sample size was 300. Further, as kurtosis increased from 1 to 7, negatively biased results were noted when the correlation among variables equaled *r* = .30, and cluster size was greater than the number of clusters. When the correlation among variables increased in normal data conditions from *r* = .50 to *r* = .70, fit bias was negative. As sample size increased, especially as the number of clusters increased, fit bias was negligible with similar results reported across estimation methods.

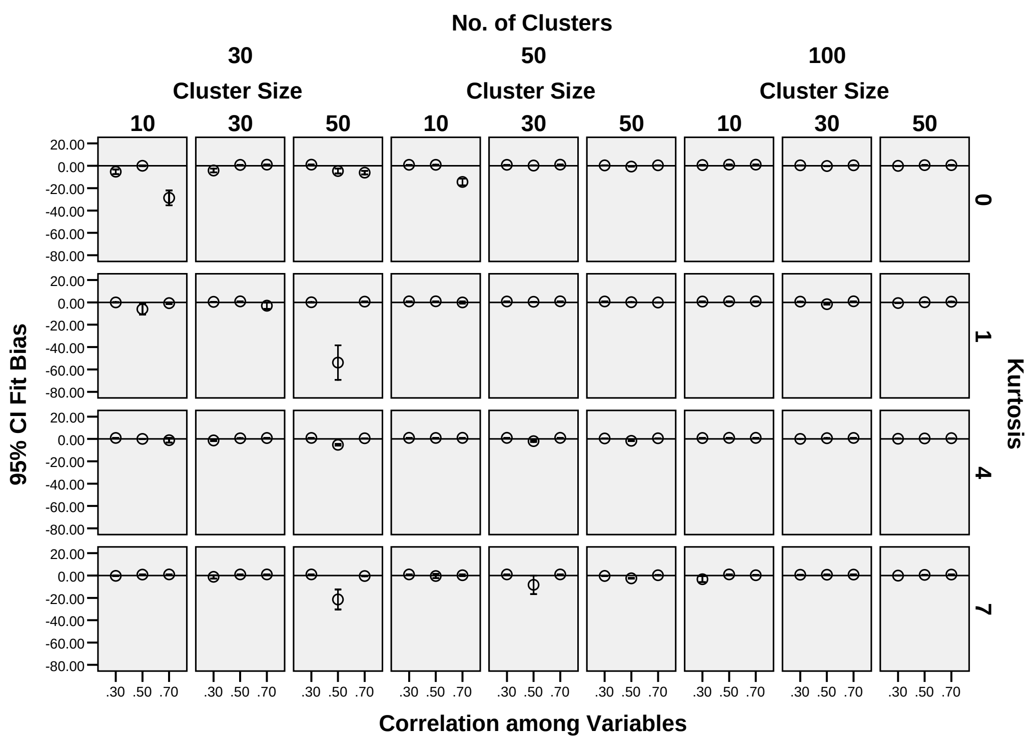

Figure 26. Confidence intervals examining fit bias by kurtosis, correlation among variables, the number of clusters, and cluster size.

The results reported in Table 78 regarding comparative fit index bias indicated that each of the design conditions as main effects was statistically significant, with the exception of skew among variables. However, similar to the results reported for the prior fit indices, each statistically significant main effect explained less than one percent of the variation in the outcome variable. Similar results were found for the interaction terms with the exception of the interaction terms that included skewness among variables, which was not statistically significant. Consistent with the prior comparative fit indices for models one and two examined in this study, each statistically significant interaction effect explained less than one percent of the variation in bias associated with outcome variable.

Regarding the post hoc analysis, the results displayed in Figure 27 revealed that when the sample size was 300 (i.e., the number of clusters was 30 and cluster size was 10) comparative fit index bias was negligible for the maximum likelihood estimator in normal data conditions. In comparison to the maximum likelihood estimator, negative CFI bias was noted for the generalized and weighted least squares estimators in this same scenario. When the correlation among variables was extreme (r = .70), comparable results were found when the number of clusters equaled 50 and cluster size was 10. Similar to the results reported earlier, as sample size increased to 900 and greater, CFI bias was negligible across estimation methods with little to no bias reported despite the magnitude of the correlation among variables.

Table 78
Factorial ANOVA Results of the Study's Six Design Conditions Effect on CFI Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	198745.7	88	2070.267	66.401	< .001	.029[a]
Intercept	11678.63	1	11678.64	374.574	< .001	.002
Est. Mthd	5634.212	2	2817.106	90.354	< .001	.001
Clstrs	11619.33	2	5809.666	186.336	< .001	.002
Clstr. Size	22499.56	2	11249.78	360.819	< .001	.003
Corr	7867.395	2	3933.697	126.167	< .001	.001
Skew	72.575	3	24.192	.776	.507	.000
Kurt	9849.441	3	2462.360	78.976	< .001	.001
Est. Mthd. * Clstrs	3918.954	4	979.739	31.424	< .001	.001
Est. Mthd. * Clstr. Size	6492.381	4	1623.095	52.058	< .001	.001
Est. Mthd. * Corr	2720.852	4	680.213	21.817	< .001	.000
Est. Mthd. * Skew	10.418	6	1.736	.056	.999	.000
Est. Mthd. * Kurt	6481.141	8	810.143	25.984	< .001	.001
Clstrs * Clstr. Size	13928.39	4	3482.099	111.683	< .001	.002
Clstrs * Corr	3739.812	4	934.953	29.987	< .001	.001
Clstrs * Skew	104.447	6	17.408	.558	.764	.000
Clstrs * Kurt	12069.04	8	1508.631	48.387	< .001	.002
Clstr. Size * Corr	9600.394	4	2400.098	76.979	< .001	.001
Clstr. Size * Skew	45.253	6	7.542	.242	.963	.000
Clstr. Size * Kurt	19120.29	8	2390.037	76.657	< .001	.003
Corr * Skew	43.347	6	7.225	.232	.966	.000
Corr * Kurt	24616.81	8	3077.102	98.693	< .001	.004
Skew * Kurt	43.134	1	43.134	1.383	.240	.000
Error	6647282	213201	31.178			
Total	6853598	213298				
Corrected Total	6846028	213297				

a. R Squared = .029 (Adjusted R Squared = .029)

Note. The model eta squared for the full factorial 6-way model was .113 with df=696. The model eta squared for only the main and the two-way interaction effects was .029 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .084 (i.e., .113-.029).

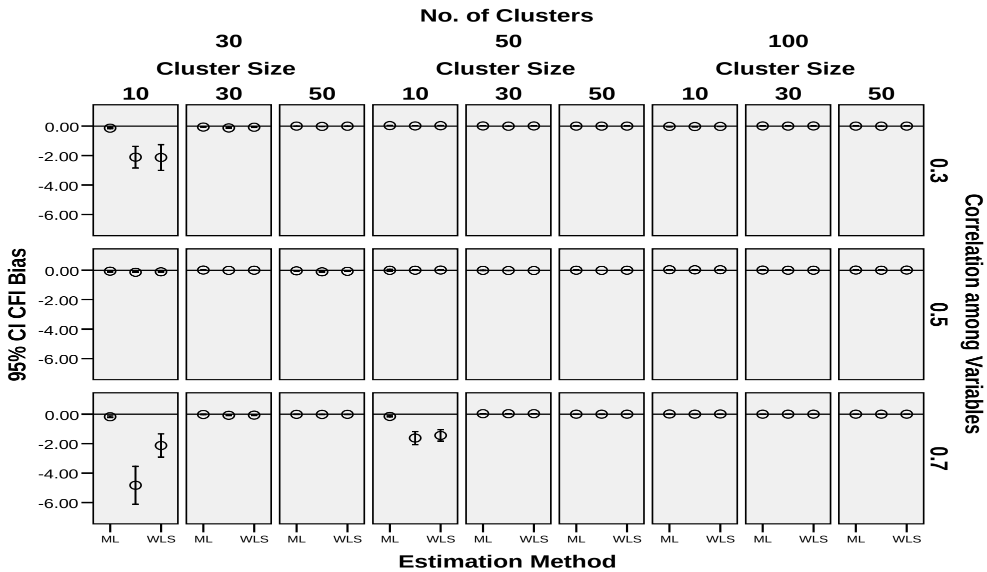

Figure 27. Confidence intervals examining CFI bias by estimation method, the number of clusters, cluster size, and the correlation among variables.

The results displayed in Table 79 reveal that each of the design conditions as main and interaction effects was a statistically significant factor impacting RMR bias. Regarding main effects, the number of clusters explained the greatest amount of variation; however, the amount of variance explained in the outcome variable was less than one percent (partial η^2=.006). As for interaction effects, cluster size by the correlation among explained approximately two percent of the variation in the dependent variable (partial η^2=.018), while the remaining terms explained less than one percent of the variance in the criterion variable. Post hoc confidence intervals (if statistically significant results) are displayed in Figure 28.

Table 79
Factorial ANOVA Results of the Study's Six Design Conditions Effect on RMR Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	6972.718	88	72.632	144.941	< .001	.061[a]
Intercept	211.186	1	211.186	421.432	< .001	.002
Est. Mthd	170.734	2	85.367	170.354	< .001	.002
Clstrs	654.852	2	327.426	653.393	< .001	.006
Clstr. Size	109.472	2	54.736	109.228	< .001	.001
Corr	498.161	2	249.081	497.051	< .001	.005
Skew	8.261	3	2.754	5.495	.001	.000
Kurt	53.549	3	13.387	26.715	< .001	.001
Est. Mthd. * Clstrs	150.242	4	37.561	74.954	< .001	.001
Est. Mthd. * Clstr. Size	110.947	4	27.737	55.350	< .001	.001
Est. Mthd. * Corr	6.189	4	1.547	3.088	.015	.000
Est. Mthd. * Skew	7.073	6	1.179	2.352	.028	.000
Est. Mthd. * Kurt	36.471	8	4.559	9.097	< .001	.000
Clstrs * Clstr. Size	1137.505	4	284.376	567.485	< .001	.011
Clstrs * Corr	971.214	4	242.804	484.525	< .001	.009
Clstrs * Skew	58.379	6	9.730	19.416	< .001	.001
Clstrs * Kurt	73.423	8	9.178	18.315	< .001	.001
Clstr. Size * Corr	1910.996	4	477.749	953.369	< .001	.018
Clstr. Size * Skew	55.269	6	9.212	18.382	< .001	.001
Clstr. Size * Kurt	75.579	8	9.447	18.853	< .001	.001
Corr * Skew	24.169	6	4.028	8.038	< .001	.000
Corr * Kurt	70.300	8	8.787	17.536	< .001	.001
Skew * Kurt	.551	1	.551	1.100	.294	.000
Error	106839.6	213203	.501			
Total	113983.9	213300				
Corrected Total	113812.3	213299				

a. R Squared = .061 (Adjusted R Squared = .061)

Note. The model eta squared for the full factorial 6-way model was .115 with df=696. The model eta squared for only the main and the two-way interaction effects was .061 with df=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .054 (i.e., .115-.061).

The results displayed in Figure 28 suggest that as the correlation among variables increased, RMR bias was negatively impacted. When cluster size was greater than the number of clusters (i.e., when the number of clusters = 30 and cluster size = 50), RMR bias was negatively impacted when the correlation among variables was held at r =.30. As sample size increased, especially when the number of clusters was 100, little to no bias was noted across estimation methods with similar results reported across estimation methods.

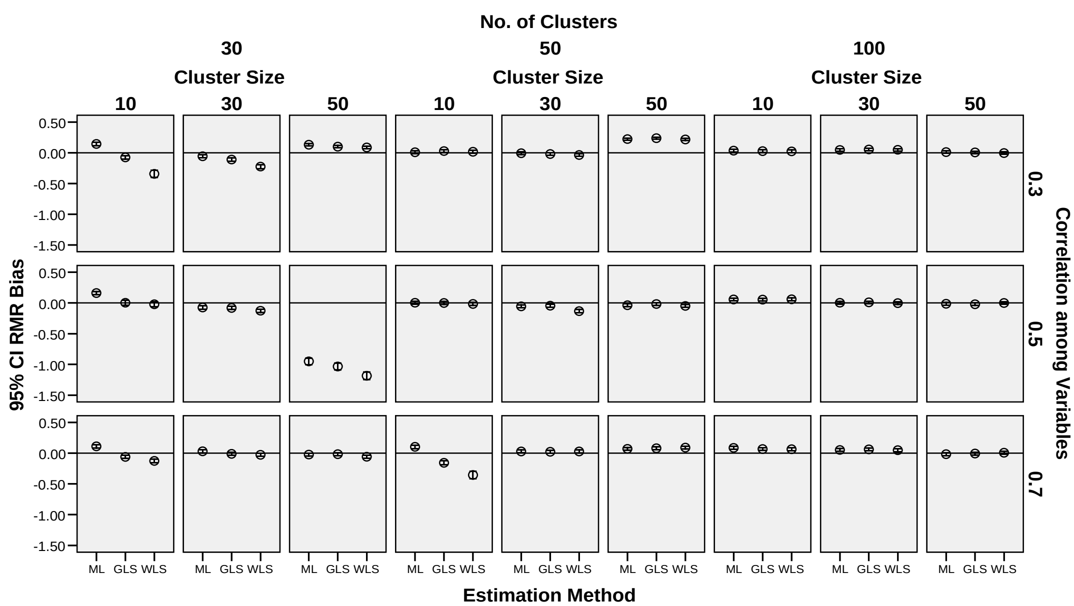

Figure 28. Confidence intervals examining RMR bias by estimation method, the number of clusters, cluster size, and correlation among variable.

The results of AIC bias reported in Table 80 revealed that each of the study design conditions as main effects had a statistically significant impact on AIC bias. However, each statistically significant main effect explained less than one percent of the variation in the outcome variable. Regarding interaction terms, the term comprised of cluster size by skew was not statistically significant, while the remaining interaction effects were statistically significant. Similar to the main effects, each of the statistically significant interaction terms explained less than one percent of the variation in AIC bias.

The post hoc confidence intervals examining AIC bias by the number of clusters, cluster size, correlation among variables and estimation method are reported in Figure 29. The results indicate that in small sample sizes (n = 300), extreme correlation among variables was associated with negatively biased AIC. Negative AIC bias was most apparent when the number of clusters was held at 30 and 50 while cluster size was 10. Negative AIC bias in this scenario was most evident for both the generalized and weighted least squares estimators, while the maximum likelihood estimator produced less biased results. As the number of clusters increased to 100, little to no bias was indicated across estimation methods despite the magnitude of the correlation among variables.

Table 80

Factorial ANOVA Results of the Study's Six Design Conditions Effect on AIC Bias

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	3318188	88	34564.46	36.972	< .001	.016[a]
Intercept	126750.0	1	126750.0	135.580	< .001	.001
Est. Mthd	23198.064	2	11599.03	12.407	< .001	.000
Clstrs	247085.9	2	123543.0	132.150	< .001	.001
Clstr. Size	176526.6	2	88263.30	94.412	< .001	.001
Corr	128779.4	2	64389.70	68.875	< .001	.001
Skew	21051.193	3	7017.064	7.506	< .001	.000
Kurt	221529.5	3	55382.38	59.241	< .001	.001
Est. Mthd. * Clstrs	18235.260	4	4558.815	4.876	.001	.000
Est. Mthd. * Clstr. Size	67462.919	4	16865.73	18.041	< .001	.000
Est. Mthd. * Corr	55879.209	4	13969.80	14.943	< .001	.000
Est. Mthd. * Skew	8150.962	6	1358.494	1.453	.190	.000
Est. Mthd. * Kurt	69980.095	8	8747.512	9.357	< .001	.000
Clstrs * Clstr. Size	90255.350	4	22563.84	24.136	< .001	.000
Clstrs * Corr	97635.495	4	24408.87	26.109	< .001	.000
Clstrs * Skew	35113.089	6	5852.182	6.260	< .001	.000
Clstrs * Kurt	384982.2	8	48122.77	51.475	< .001	.002
Clstr. Size * Corr	235380.3	4	58845.07	62.945	< .001	.001
Clstr. Size * Skew	9327.579	6	1554.596	1.663	.126	.000
Clstr. Size * Kurt	273464.1	8	34183.01	36.564	< .001	.001
Corr * Skew	21340.935	6	3556.822	3.805	.001	.000
Corr * Kurt	518680.7	8	64835.08	69.352	< .001	.003
Skew * Kurt	8482.591	1	8482.591	9.074	.003	.000
Error	2.0E+008	213203	934.872			
Total	2.0E+008	213300				
Corrected Total	2.0E+008	213299				

a. R Squared = .016 (Adjusted R Squared = .016)

<u>Note</u>. The model eta squared for the full factorial 6-way model was .058 with <u>df</u>=696. The model eta squared for only the main and the two-way interaction effects was .016 with <u>df</u>=88. Thus, the eta squared effect size for all the unreported two-, three-, four-, five-, and the six-way interaction effects was .042 (i.e., .058-.016).

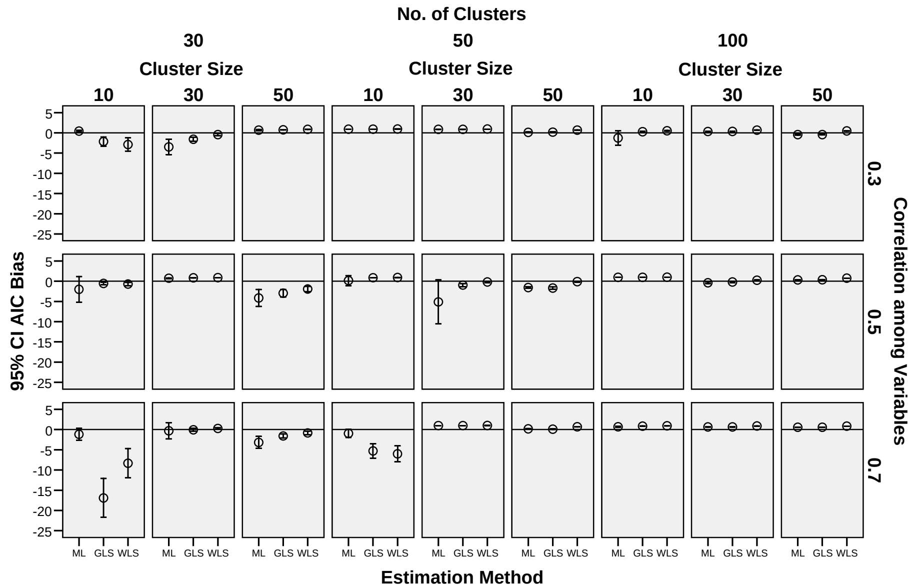

Figure 29. Confidence intervals examining AIC bias by estimation method, the number of clusters, cluster size, and kurtosis.

Summary of Model Three

Similar to the results from models one and two, the addition of the level-two parameter in model three indicated that of the six conditions examined, sample size appeared to have the greatest impact on the goodness of fit index. In model three, the addition of the level-two parameter resulted in negative GFI bias that was greater in magnitude when compared to both model one and two. Kurtosis and the correlation among variables had a statistically significant negative effect on fit index bias. Namely, as the correlation increased from r = .30 to r = .70 in small sample sizes in normal data conditions, fit index bias was negative. As kurtosis increased from one to seven, and cluster size was greater than the number of clusters (i.e., number of clusters = 30 and cluster size = 50), fit bias was negative when the correlation among variables was r = .50. Similar to model one, the AIC index was negatively biased in small sample sizes. This was most evident when the number of clusters ranged from 30 to 50 and cluster size was held at 10. Further, negative AIC bias increased in these same sample sizes as the correlation among variables increased from r = .30 to r = .70.

Sample size, estimation method and the correlation among variables appeared to be the salient factors impacting CFI bias. When the number of clusters was 30 and cluster size was 10, negative bias was noted for the generalized least squares estimator when the correlation among variables was held at r = .30. As the correlation among variables increased to r = .70, increased negative bias was noted when the number of clusters was 30 and cluster size was 10. In addition, negative CFI bias was detected when the number of cluster was 50 and cluster size was 10, with the maximum likelihood estimator producing negligible bias while severe negative bias was noted for the generalized and weighted least squares estimators.

The results of the size study conditions regarding bias among the root mean square residual revealed that the number of clusters by cluster size had the greatest impact on RMR bias (partial η^2 =.006). As the correlation among variables increased from r = .30 to r = .70, the variation in RMR bias increased in small and moderately small sample sizes. As sample size increased, RMR bias decreased across estimation methods with stable results reported for 100 clusters across all clusters sizes regardless of the amount of correlation among variables.

CHAPTER V

DISCUSSION AND CONCLUSIONS

The purpose of the present study was to compare three estimation methods in multilevel SEM under varying conditions of data normality across different sample sizes to determine which estimator produced more robust parameter estimates, standard errors, and fit indices.

This chapter summarizes the findings of the study with regard to the research questions posed in Chapter I. Relevant implications and conclusions are drawn based on these findings in terms of their potential influence on research practice. Additionally, limitations are discussed. Finally, this chapter presents recommendations for future research. The discussion is organized around the original research questions.

Research Questions

Research Question 1

> Do the parameter estimates differ by the design factors (estimation method, nonnormality conditions, the number of clusters, and cluster size)?

The results from the factorial analysis with regression parameter estimate bias as the dependent variable indicated that in each of the models investigated, *the regression coefficients were estimated with little to no bias*. Although the number of clusters, cluster size, correlation among variables, skew and kurtosis were statistically significant factors in predicting bias, each explained less than one percent of the variation in the outcome variable. Note that the estimation method was not a statistically significant factor impacting regression parameter estimate bias among the three models investigated in the study.

The results obtained from research question one were similar to the results found in prior studies, namely Brown and Draper (2000), Willson and Zhang (2003), Mass and Hox (2004), Mass and Hox (2005), and Zhang and Willson (2006). These researchers reported that the regression coefficients are estimated without bias, while their standard errors tend to be biased downward in small sample sizes, especially at the group level. In the current study, unbiased regression coefficients were prevalent in each of the three models examined, regardless of model complexity.

Table 81 provides a comparison of the results obtained by Zhang (2005) and the present study. Zhang, whose study was the most recent study closely related to the topic of the present study, found that sample size did impact the level-1 parameter estimate and explained approximately 50% of the variance in the level-1 parameter estimate’s bias (partial η^2=,488). In comparison, the present study found that the sample size explained approximately 8% of the variance in the outcome variable. While the findings from the present study were in-line with prior studies, the resulting impact of sample size on bias among the level-1 parameter estimate was much greater in Zhang's study. A plausible explanation could be attributed to the number of sample size conditions investigated. Zhang's study investigated four varying sample sizes (i.e., 10 clusters with cluster size fixed at 50, 30 clusters with cluster sized fixed at 10 and 50, and 50 clusters with cluster sized fixed at 50).The present study included 30, 50, and 100 clusters with cluster size (level-1) fixed at 10, 30 and 50 in each cluster (level-2) for a total of nine varying sample sizes. Nonetheless, further investigation is warranted to determine the effect of sample size on bias among the level-1 parameter estimate.

Regarding data normality, both studies found similar results. Zhang reported that data normality conditions explained approximately 6% of the variance in the level-1 parameter estimate, while the present study found that data normality conditions explained approximately 4% of the bias variance in the level-1 parameter estimate.

Concerning bias among the cross-level interaction estimate, similar results were found across the two studies. Namely, data normality conditions and sample size explained a negligible portion of variance in bias among the cross-level interaction term.

Table 81
Comparison of Present Study to Zhang's (2005) Study Regarding Parameter Estimate Bias

	Byrd	Zhang
	Partial η^2	Partial η^2
Level-1 Parameter Estimate		
Data Normality	.039	0.062
Sample Size	.078	0.488
Cross-level Interaction		
Data Normality	.009	0.000
Sample Size	.005	0.010

Research Question 2

Do standard errors differ by the design factors (estimation methods, nonnormality conditions, the number of cluster and cluster size)?

The number of clusters (group level) appeared to have the greatest impact on bias among the standard errors. Regarding the results of the factorial ANOVAs for the level-1 standard errors, the number of clusters explained approximately 5% of the variation in the outcome variable. However, the number of clusters explained approximately 10% of the variance in bias associated with the level-2 standard errors. When examining interaction effects via confidence intervals as a post hoc procedure, bias was most prevalent when the number of clusters and cluster size was smaller. Most notably, when the number of clusters was 30 and cluster size was held at 10, the level-1 standard errors were biased downward by approximately 20% for the maximum likelihood and generalized least squares estimators, while the weighted least squares estimator produced level-1 standard errors that were negatively biased by 25% when the correlation among variables was fixed at .7. Regarding the level-2 standard errors in model one, the level-2 standard errors were biased downward by approximately 24% across each of the estimation methods examined in nonnormal data when the correlation among variables was fixed at .5 and kurtosis was held constant at 7. In this same setting (30 clusters with cluster size fixed at 10), when kurtosis was fixed at 4 and the correlation among variables was held at .7, both the maximum likelihood and generalized least squares estimators resulted in standard errors that were biased downward by approximately 11%, while the weighted least squares estimator resulted in standard errors that were negatively biased by approximately 8%. As the number of clusters increased from 30 to 50 and cluster size was fixed at 10, the amount of negative bias associated with the level-2 standard error decreased to approximately 2% for each estimation method in normally distributed data. When the number of clusters was 100 and cluster size was fixed at 30 and 50 respectively, bias associated with the level-2 standard errors was near zero, with each estimator producing similar results. Similar results regarding standard error bias were found for each of the models investigated in the study.

The results reported for research question two were in line with prior studies which investigated bias among standard errors in the HLM context. Mass and Hox (2005) found that standard errors for the regression coefficients were slightly biased downward when the number of groups was less than 50. Regarding the individual sample size, in Monte Carlo simulation studies, Willson and Zhang (2003) and Zhang and Willson (2006) reported that the first level sample size did matter. In a two-level cross-interaction model, power did not increase further when the group size exceeded 35, when the number of groups was fixed at 120. The underlying message is that the regression coefficients are unbiased while standard errors are biased downward in small sample sizes, especially when the number of clusters is 50 or less and cluster size is less than 30.

Empirical researchers must pay attention to sample size and data normality conditions. Regarding sample size, it is recommended that researchers strive to maintain a level-1 sample size of at least 30 with a minimum of 30 groups at level-2 (group-level) based on the results of this study. The results of this study regarding the impact of the number of groups or clusters and the impact on standard error bias further underscore the findings from earlier studies examining multilevel regression. For example, Kreft (1996) indicated that researchers should strive for a sample of at least 30 groups with at least 30 individuals per group if interest is primarily in the fixed effects. Hox (n.d.) suggested that sample size should be increased to 50/20 (50 groups with a minimum of 20 individuals per group) if interest is on the cross-level interactions. Further, if interest lies in the random portion of the model, Hox suggested that the variance and covariance should be based on a 100/10 rule (100 groups with a minimum of 10 individuals per group).

As data depart from normality, especially when kurtosis is severe, researchers should consider the weighted least squares estimation method in moderately large sample sizes. However, in large sample sizes, especially when the number of clusters is at least 100 and cluster size is fixed at 30 or greater, Browne (1884) and Ollson, Foss, Troye, and Howel (2000) have shown that different estimation procedures such as ML, GLS, and the asymptotically distribution free weighted least squares estimator (WLS), will produce estimates that converge and possess similar asymptotic properties in factor analysis. The results from the present study revealed that in large sample size, especially when the number of cluster was 100 and cluster size was at least 30, different estimation methods (namely maximum likelihood, generalized least squares, and weighted least squares) produce similar results, further indicating that the findings from Brown and Ollson, et al. apply to multilevel SEM.

Similar to research question one, Table 82 presents a comparison of the present study’s findings to Zhang’s results. Regarding the level-1 and cross-level interaction standard errors, similar results were found across studies. Namely, data normality conditions explained little variation in the outcome variable. Zhang reported that approximately 7% of the variance among the levl-1 and interaction standard errors was explained by the data normality conditions investigated. Similarly, the present study revealed that data normality conditions explained approximately 1% of the variance in the outcome variables. Note that 9 varying degrees of data normality were investigated in the present study, while only three data normality conditions were included in Zhang’s study (i.e., normal[skew and kurtosis=0],moderately nonnormal [skew = .5, kurtosis = 1], and severely nonnormal [skew = 1, kurtosis = 3.75]).

Sample size was found to explain an increasing amount of bias in the level-1 and cross-level interaction terms in both studies. Zhang found that over 90% of the variance in the outcome variables were explained by sample size, while the present study found that approximately 14% of bias variance was accounted for by the total sample size (i.e., number of clusters by cluster size). In summary, both studies were in agreement that sample size had the greatest impact on the level-1 and cross-level standard errors.

Table 82
Comparison of Present Study to Zhang's (2005) Study Regarding Standard Error Bias

	Byrd	Zhang
	Partial η^2	Partial η^2
Level-1 Standard Error		
Data Normality	.007	0.068
Sample Size	.136	0.912
Cross-level Standard Error		
Data Normality	.007	0.061
Sample Size	.133	0.923

Research Question 3

Do fit indices differ by the design factors (estimation methods, nonnormality conditions, and sample and cluster size)?

Among the six study design conditions examined, sample size at levels-1 and -2 and data nonnormality appeared to have the greatest impact on the goodness of fit index. In each of the three models considered in this study, especially when the number of clusters ranged from 30 to 50 and cluster size was fixed at 10, the GFI was negatively biased for each of the estimation methods examined. In more succinct terms, the amount of negative bias associated with the maximum likelihood and generalized least squares estimators was approximately 5% while the weighted least squares estimation method resulted in a GFI that was negatively biased by 10% in normal data conditions. As the data departed from normality in these same sample sizes, negative bias increased for both the generalized and weighted least squares estimators. Namely, in moderately nonnormal data conditions, the amount of negative bias associated with the maximum likelihood and generalized least squares estimators was approximately 18%, while the weighted least squares estimator produced a GFI that was negatively biased by approximately 22%. Note that the maximum likelihood estimator produced GFI estimates that were less than 5% in each of the models investigated when compared to both the generalized and maximum likelihood estimation methods. The result of the factorial ANOVA analyses revealed that the number of clusters explained approximately 14% of the variation in the outcome variable while cluster size explained approximately 8% of the variance in GFI bias. The results were similar to those reported by Fan et al. (1999) and Hu and Bentler (1995, 1998, 1999) who concluded that the GFI is sample size dependent.

Kurtosis and the correlation among variables had a statistically significant negative effect on fit index bias. Namely, as the correlation increased from r = .30 to r = .70 in small sample sizes and in normal data conditions, fit index bias was negatively biased by approximately 20%. As kurtosis increased from 1 to 7, and cluster size was greater than the number of clusters (i.e., number of clusters = 30 and cluster size = 50), fit bias was negative and ranged from -5% when kurtosis ranged from

1 to 4 to -60% when kurtosis was held at 7 and the correlation among variables was *r* = .50. In addition, the AIC index was negatively biased in small sample sizes in the more complex Model 3, while little bias was noted in models 1 and 2. This was most evident in model 3 when the number of clusters ranged from 30 to 50 and cluster size was fixed at 10. In this scenario, AIC bias was negligible for the maximum likelihood estimator (< 5%), while 12% negative bias was noted for the generalized lest squares estimator when the correlation among variables was *r* = .30. When the correlation among variables increased to *r* = .70, negative bias (18%) was noted for the generalized least squares estimator while 3% negative bias was associated with the weighted least squares estimator. Similar results were noted when the number of clusters was fixed at 50 and cluster size was 10. Further, negative AIC bias increased in these same sample sizes as the correlation among variables increased from *r* = .30 to *r* = .70.

Sample size, estimation method, and the correlation among variables appeared to be the salient factors impacting CFI bias. When the number of clusters was 30 and cluster size was fixed at 10, negative bias was noted for the generalized least squares estimator when the correlation among variables was held at *r* = .30. As the correlation among variables increased to *r* = .70, increased negative bias was noted (ranging from -2% to approximately

-20%) when the number of clusters was 30 and cluster size was 10. In addition, negative CFI bias was detected when the number of cluster was 50 and cluster size was 10, with the maximum likelihood estimator producing negligible bias while 22% negative bias was noted for the generalized least squares estimators and approximately 24% negative bias reported for the weighted least squares estimators.

The results of the six study conditions regarding bias among the root mean square residual revealed that the number of clusters by cluster size had the greatest impact on RMR bias (partial η^2 =.006). As the correlation among variables increased from r = .30 to r = .70, the variation in RMR bias increased in small and moderately small sample sizes in Model 3. As sample size increased, RMR bias decreased from 30% in small sample sizes to less than 5% bias across estimation methods with stable results reported for 100 clusters across all clusters sizes regardless of the amount of correlation among variables. In Models 1 and 2, kurtosis was a significant factor. Similar to the results reported for Model 3, as kurtosis increased from 1 to 7, negative bias (~40%) was associated with smaller sample sizes, especially when cluster size ranged from 10 to 30 and the number of clusters ranged from 30 to 50.

In summary, the performance of GLS and ML with respect to empirical fit was reasonably robust to moderate deviations from multivariate normality in both the moderate to large sample sizes investigated in the study. The findings from the current study regarding the performance of the ML and GLS estimation methods compared favorably to earlier studies conducted by Chou, Bentler, and Satorra (1991). However, fit indices associated with the weighted least squares estimator were negatively biased in small sample sizes. Zhang's study did not investigate fit indices. Therefore, no comparisons were included in the present study.

Conclusion

Violations of normality and performance of various estimators have been broadly explored in single-level data analysis (e.g., Bickel, 2007; Bollen, 1989; Brown, 2006; Finely & DiStefano, 2006; Olsson, Foss & Howell, 2000). The present study studied these topics from the perspective of a *multilevel data structure and with SEM analysis methods*. The results revealed that the regression coefficients are estimated with little bias among the study design conditions investigated. However, the number of clusters (group level) appeared to have the greatest impact on bias among the standard errors. In small sample sizes (i.e., 300 and 500) the standard errors were negatively biased by more than 25%. Further, negative standard error bias increased in magnitude as the correlation among variables and kurtosis increased. This was expected as earlier studies have shown that nonnormality, especially high kurtosis, can produce poor estimates and grossly incorrect standard errors and hypothesis tests, even in large samples. However, in the current study, as the data exhibited departure from normality in moderate to large sample sizes, the weighted least squares estimator produced slightly less biased standard errors (an improvement of approximately 10%) when compared to the maximum likelihood and generalized least squares estimators.

Regarding fit statistics, negative bias was noted among each of the fit indices investigated when the number of clusters ranged from 30 to 50 and cluster size was fixed at 10. The least amount of bias was associated with the maximum likelihood estimator in each of the data normality conditions examined. As sample size increased, bias decreased sharply to near zero (< 5%) with similar results reported across estimation methods. Note that the number of predictors at both level-1 and 2 impacted AIC and RMR bias with increased bias noted in the more complex Model 3 in small to moderately small sample sizes.

Recommendations

Empirical researchers must pay attention to sample size and data normality conditions. Regarding sample size, it is recommended that researchers strive to maintain a level-1 sample size of at least 25 to 30 and at least 30 clusters in multilevel SEM. In addition, as data exhibit departure from normality, especially when kurtosis is severe, researchers should consider the weighted least squares estimation method in moderately large sample sizes. However, in large sample sizes, especially when the number of clusters is at least 100 and cluster size is near 30, similar results should be obtained regardless of estimation method. Nevertheless, in small sample sizes (below 1,000) it is recommended that of the three estimation methods investigated in this study, empirical researchers employ the maximum likelihood estimator. The rationale for recommending the maximum likelihood estimator is based on the least amount of bias associated with both the standard errors and fit indices. Among the fit indices investigated in this study, the least amount of bias (< 5%) was associated with the maximum likelihood estimator in each of the six study design conditions examined.

Limitations

Due to practical constraints, the present study only investigated multilevel data in three simple regression models. Future research should replicate this study utilizing more complex models. In addition, this study only examined the maximum likelihood, generalized least squares, and weighted least squares estimators. Future studies should include the robust maximum likelihood estimator in addition to the three estimation methods examined in this study as the robust maximum likelihood estimator is commonly employed in HLM analysis. Finally, it would also be meaningful to include additional statistical fit indices as this is one of the advantages for conducting multilevel regression analysis in the SEM framework.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csake (eds.), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267-281.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.*, 30, 9-14.

Arbuckle, J. L., & Wothke, W. W. (1999). *Amos 4.0 user's guide*. Chicago: SmallWaters.

Barcikowski,R.S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6*, 267-285.

Bauer,D.J. (2003). Estimating multilevel linear models as structural models. *Journal of Educational and Behavioral Statistics, 28,* 135-167.

Bentler, P.M. (1989). *EQS structural equations program manual*. Los Angeles: BMDP Statistical Software.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P.M. & Liang, J. (2003). Tow-level mean and covariance structures: Maximum liklihhod via EM algorithm. In S.P. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 53-70). Hillsdale, NJ: Erlbaum.

Bentler, P. M., & Wu, E. J. C. (2002). EQS 6 for Windows user's guide. Encino, CA: Multivariate Software, Inc.

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* The Guilford Press: New York.

Bollen, K.A.(1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K.A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107*, 256-259.

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited.In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation models: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 139–168). Chicago: Scientific Software International.

Bozdogan, H. (1987). Model-selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Brown, T. (2006). *Confirmatory factor analysis for applied research*. The Gilford Press: New York.

Browne, M.W. (1984). Asymptotic distribution-free methods in the analysis of covariate structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62-83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Browne, W. J. & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics, 15*, 391-420.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Burnham, K. P. & Anderson, D. R. (1998*). Model selection and inference: A practical information-theoretic approach*. New York: Springer-Verlag.

Byrd, J.K. (2007). From the field: A call for statistical reform in EAQ. *Educational Administration Quarterly, 43,* 381-391.

Chou, C. P.,&Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Issues and applications* (pp. 37-55). Newbury Park, CA: Sage.

Chou, C.P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44*, 347-357.

Cummings, G. (2007).Inference by eye: Pictures of confidence intervals and thinking about levels of confidence.*Teaching Statistics,29,* 89-92.

Curran, P. J., West, S. G, & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16-29.

Curran, P. (2003). Have multiplevel models been structural equation models all along? *Multivariate Behavioral Research, 34*, 529-569.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Education and Behavioral Statistics, 20*, 115-147

Dempster, A., Lair, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39, 138.

Bryk,A.S., Raudenbush, S.W., Seltzer, M. & Condon, R. (1988). *An introduction to HLM: Computer program and user's guide*. Chicago: University of Chicago, Department of Education.

Fan, X, Thompson, B. & Wang, L. (1999). Effects of ample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6*, 56-83.

Eliason, S.R.(1993). *Maximum likelihood estimation.* Sage: Newbury Park, Ca.

Finney, S. & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. IN Hancock, G, & Mueller, R. *Structural equation modeling: A second course*, (pp.269-314). Greenwich, CT: Information Age Publishing.

Gagne, P. (n.d.). Best regression using information criteria. Retrieved July 24, 2007; from http://www.education.umd.edu/EDMS/Latent/BestRegression.doc

Gibson, N. (2003). Treatment of missing data at the second level of hierarchical linear models. *Educational and Psychological Measurement, 63*, 204-238.

Goldstein, H. (1995). *Multilevel statistical models*. London: Edward Arnold.

Goldstein, H., & McDonald, R.P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53*, 455-467.

Goldstein, H., Browne, W. J, & Rasbash, J. (2002). Partitioning variation in multilevel models. *U*nderstanding Statistics, 1, 223-232.

Hancock, G. R., & Mueller, R. O. (Eds.). (2006). *Structural equation modeling: A second course*. Greenwich, CT: Information Age Publishing, Inc.

Heck, R. & Thomas, S. (2000). *An introduction to multilevel modeling techniques*. Mahwah, NJ: Lawrence Erlbaum.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329-367.

Hoogland, J.J. (1999). *The robustness of estimation methods for covariance structure analysis*. Unpublished doctoral dissertation, University of Groningen, The Netherlands.

Hox, J.J. (n.d) Multilevel modeling: When and why? In Balderjahn, I. , Mathar, R. & Schader, M. (Eds.). *Classification, data analysis, and data highways* (pp.147-154). New York: Springer Verlag.

Hox, J.J. (1995). *Applied multilevel analysis*. Amsterdam: TT-Publikaties

Hox J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Hox, J.J. & Kreft, I.G. (1994). Multilevel analysis methods. *Sociological Methods & Research, 22*, 283-299.

Hox, J.J. & Maas, C.J.M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling 8*, 157-174.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*, 351-362.

Hu, L., & Bentler, P. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling. Concepts, issues, and applications* (pp. 76-99). London: Sage.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure analysis: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424-453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-•55.

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL : A guide to the program and applications (2nd ed.). New York: SPSS.*

Jöreskog, K. G., & Sörbom, D. (1989). LISREL 7 user's reference guide. Chicago: SPSS Publications.

Jöreskog, K. G., & Sörbom, D. (1993). *Structural equation modeling with the SIMPLIS command language*. Chicago: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 user's reference guide*. Chicago: Scientific Software International.

Jöreskog, K. G., Sörbom, D., Du Toit, S. H. C., & Du Toit, M. (2000). *LISREL 8: New statistical features*. Chicago: Scientific Software International.

Kaplan, D. (2000). *Structural equation modeling: Foundation and extensions*. Thousand Oaks, CA: Sage Publications.

Kish, L.(1965). *Survey sampling.* New York: John Wiley & sons.

Kreft, I., & de Leeuw, J. & van der Leeden, R. (1994). Review of five multilevel analysis programs: BMPD-5v, GENMOD, HLM, ML3, VARCL. *The American Statistician,48*, 324-335.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.

Kreft, I.G.G. (1996). Are multilevel techniques necessary? An overview, including simulation studies. California State University Los Angeles. Retrieved on May 29, 2007; from http://www.calstatela.edu/faculty/ikreft/quarterly/quarterly.html.

Laird, N.M. & Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics, 38*, 963-974.

Lei, M. & Lomax, R. (2005). The effect of varying nonnormality in structural equation modeling. *Structural Equation Modeling, 1,* 1-27.

Longford, N. T. (1989). Fisher scoring algorithm for variance component analysis of data with multilevel structure. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 311–317). San Diego, CA: Academic Press.

Maas. C.J.M.& Hox, J.J. (2005). Sufficient sample size for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral & Social Sciences, 1,* 86-92.

Maas. C.J.M.& Hox, J.J. (2004a).The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational statistics and data analysis, 46*, 427-440.

Maas. C.J.M.& Hox, J.J. (2004b). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58,* 127-133.

McArdle, J.J. & Hamagami, F. (1996). Multilevel models from a multiple froup structural equation perspective. In G. Marcoulides and R. Schumacker (Eds), *Advanced structural equation modeling techniques* (pp. 89-124). Hillsdale, NJ:Lawrence Erlbaum Associates.

McQuarrie, A. D. R. & Tsai, C.-L. (1998). *Regression and Time Series Model Selection.* Singapore: World Scientific.

Mehta, P.D. & Neale, M.C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*, 259-284.

Meredith, W. & Tisak, J. (1984, July). "*Tuckerizing" growth models*. Paper presented at the annual meeting of the Psychometric Society, Santa Barbara, CA.

Meredith, W. & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107-122.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 3-12.

Muthén, B. O., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19–30.

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171–189.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4*, 599-620.

Muthén, B. O., & Muthén, L. (2001). *MPLUS user's guide* (2nd ed.). Los Angeles: Author.

Muthén, L. K., & Muthén, B. O.(1998-2006). *Mplus user's guide (4th ed.)* Los Angeles, CA: Muthén & Muthén.

Muthen, B.O. & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25,* 267-316.

National Center for Education Statistics. (NCES, 2004). *User's guide to the longitudinal kindergarten–third grade public-use data file*. NCES 2004-088.

Newsome, J.T. (2002). A multilevel structural equation model for dyadic data. *Structural Equation Modeling, 9*, 431-447.

Nezlek, J. B. (2003). Using multilevel random coefficient modeling to analyze social interaction diary data. *Journal of Social and Personal Relationships, 20*, 437-469.

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event and interval contingent data in social and personality psychology research. P*ersonality and Social Psychology Bulletin, 27*, 771-785.

Olsson, U.H., Foss, T., Troye, S.V., & Howell, R.D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonormality. *Structural Equation Modeling, 7*, 557-595.

Rasbash J., Browne W., Goldstein H., Yang M., Plewis I., Healy M., Woodhouse G., Draper D., Langford I., Lewis T. (2000) *A users guide to MLwiN* , Version 2.1. London, Institute of Education.

Raudenbush ,S.W. & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd. Ed.) Newbury Park, CA: Sage.

Roberts, J.K. (2004). An introductory primer on multilevel and hierarchical linear modeling. *Learning Disabilities: A Contemporary Journal 2*, 30-38.

Rovine, M.J. & Molenaar, P.C. (1998). A nonstandard method for estimating a linear growth model in LISREL. International *Journal of Behavioral Research, 22*, 453-473.

Rovine, M.J. & Molenaar, P.C. (2000). A structural equation model to a multilevel random coefficients model. *Multivariate Behavioral Research, 35*, 51-88.

Rovine, M.J. & Molenaar, P.C. (2001). A structural equation modeling approach to the general linear mixed model. In L.M. Collins & A.G. Sayer (Eds.), *New methods for the analysis of change* (pp. 65-96). Washington, DC: American Psychological Association.

SAS (1999). *The calis procedure*. Retrieved from the following website on June 21, 2007; from http://www.tau.ac.il/cc/pages/docs/sas8/stat/chap19/sect1.htm

SAS Institute. (2002). Statistical analysis system (Version 8.1). Cary, NC: SAS Institute.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models:Ttests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*, 23-74. Retrieved on June 22, 2007; from http://www.mpr-online.de/issue20/art2/mpr130_13.pdf.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Scott L. J., Ervin, L.H.(2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician, 54*, 217-224.

Singer, J.(1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Education and Behavioral Statistics*, *2*, 323-355.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173-180.

Steiger, J. H. (1995). Structural equation modeling (SEPATH). In *STATISTICA 5, Vol. 3* [Computer software and manual] (pp. 3539–3688). Tulsa, OK: StatSoft.

Steiger, J. H. (2001). Driving fast in reverse: The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association*, *96,* 331–338.

Tanaka, J. S., & Huba, G. J. (1984). Confirmatory hierarchical factor analyses of psychological distress measures. *Journal of Personality and Social Psychology, 46*, 621-635.

Tomarken, A.J. & Baker, T. (2003). Introduction to the Special Section on Structural Equation Modeling. *Journal of Abnormal Psychology, 112*, 523-525.

Tomarken, A. J. & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology 1*, 31-65.

Van Der Leeded, R., Busing, F. & Meijer, E. (1997). *Applications of bootstrap methods for two-level models.* Unpublished paper, Multilevel Conference, Amsterdam, April 1-2, 1997.

Van Der Leeded, R., Busing, F. (1994). *First interation versus IGLS/RIGLS estimates in two-level models: A Monte Carlo study study with Ml3*, unpublished manuscript, Department of Psychometrics and Research Methodology, Leiden University.

West, S. G, Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56-75). Thousand Oaks, CA: Sage.

Willson, V. L. & Zhang, D. (2003, April). Comparing HLM, HLM-SEM, and Residual SEM Analyses for Multilevel Models. Paper presented at the 84th Annual Meeting of the American Educational Research Association, Chicago, Ill.

Zhang, D. (2005). *A Monte Carlo investigation of robustness to nonnormal incomplete data of multilevel modeling.* Unpublished doctoral dissertation, Texas A&M University - College Station.

Zhang, D., & Willson V. L. (2006). Comparing empirical power of multilevel structural equation models and hierarchical linear models: Understanding Cross-level Interactions. *Structural Equation Modeling,13*, 615-630.

APPENDIX

Syntax to Generate Fleishman Coefficients for the Desired Skewness and Kurtosis

```
       PROC IML;
SKEWKURT={-1 2.5,
           2 7,
           0 0,
          -2 7};

START NEWTON;
  RUN FUN;
  DO ITER = 1 TO MAXITER
  WHILE(MAX(ABS(F))>CONVERGE);
        RUN DERIV;
        DELTA=-SOLVE(J,F);
        COEF=COEF+DELTA;
        RUN FUN;
  END;
FINISH NEWTON;
MAXITER=25;
CONVERGE=.000001;
START FUN;
  X1=COEF[1];
  X2=COEF[2];
  X3=COEF[3];
  F=(X1**2+6*X1*X3+2*X2**2+15*X3**2-1)//
    (2*X2*(X1**2+24*X1*X3+105*X3**2+2)-SKEWNESS)//
    (24*(X1*X3+X2**2*(1+X1**2+28*X1*X3)+X3**2*
      (12+48*X1*X3+141*X2**2+225*X3**2))-KURTOSIS);
FINISH FUN;
START DERIV;
  J=((2*X1+6*X3)||(4*X2)||(6*X1+30*X3))//
    ((4*X2*(X1+12*X3))||(2*(X1**2+24*X1*X3+105*X3**2+2))
     ||(4*X2*(12*X1+105*X3)))//
    ((24*(X3+X2**2*(2*X1+28*X3)+48*X3**3))||
     (48*X2*(1+X1**2+28*X1*X3+141*X3**2))||
     (24*(X1+28*X1*X2**2+2*X3*(12+48*X1*X3+141*X2**2+225*X3**2)

     +X3**2*(48*X1+450*X3))));
FINISH DERIV;
DO;
NUM = NROW(SKEWKURT);
DO VAR=1 TO NUM;
  SKEWNESS=SKEWKURT[VAR,1];
  KURTOSIS=SKEWKURT[VAR,2];
  COEF={1.0, 0.0, 0.0};
  RUN NEWTON;
  COEF=COEF`;
  SK_KUR=SKEWKURT[VAR,];
  COMBINE=SK_KUR || COEF;
  IF VAR=1 THEN RESULT=COMBINE;
  ELSE IF VAR>1 THEN RESULT=RESULT // COMBINE;
END;
  PRINT "COEFFICEINTS OF B, C, D FOR FLEISHMAN'S POWER TRANSFORMATION";
  PRINT "Y = A + BX + CX^2 + DX^3";
  PRINT " A = -C";
  MATTRIB RESULT COLNAME=({SKEWNESS KURTOSIS B C D})
                 FORMAT=12.9;
  PRINT RESULT;
END;
QUIT;
```

Syntax to Generate Intermediate Pair-wise Correlation Coefficients

```
DATA D1;
  B1=.978350485; C1=-.124833577; D1=.001976943;  * use Fleishman
coefficients;
  B2=.978350485; C2= .124833577; D2=.001976943;
  TARGET=.70;                                    * target population
correlation;
  R=.5;                                          * starting value for
iteration;

DO I=1 TO 5;

FUNCTION=(R**3*6*D1*D2+R**2*2*C1*C2+R*(B1*B2+3*B1*D2+3*D1*B2+9*D1*D2)-
TARGET);
   DERIV=(3*R**2*6*D1*D2+2*R*2*C1*C2+(B1*B2+3*B1*D2+3*D1*B2+9*D1*D2));
   RATIO=FUNCTION/DERIV;
   R_TEMP = R - RATIO;
   IF ABS(R_TEMP - R)>.00001 THEN R = R_TEMP; OUTPUT;
END;
PROC PRINT; WHERE I=5;    * print intermediate correlation r for the
last iteration;
  VAR I RATIO R;
RUN;
```

VITA

Name:	Jimmy Kent Byrd
Address:	PO Box 2998, Denton, Texas 76202
Email Address:	jbyrd@coe.unt.edu

Education:	B. S., Agriculture Science,
	Tarleton State University, 1989
	M. Ed. Educational Administration,
	Tarleton State Univesity, 1995
	Ph. D., Educational Administration,
	Texas A&M University, 2001
	Ph. D., Educational Psychology,
	Texas A&M University, 2008